PCP and conformational motifs: applications in predicting mechanical strength, allergenicity of proteins and vaccine design

Wenzhe Lu

University of Texas Medical Branch

Thesis Dissertation

Table of Contents

Table of Contents **2**

List of Figures **5**

List of Tables **6**

List of Abbreviations **7**

PREFACE **8**

DEDICATION 8

THANKS 8

ABSTRACT **10**

CHAPTER 1: INTRODUCTION **12**

1.1 Protein sequence motifs 12

1.2 Physical–Chemical Properties (PCP) of amino acids 13

1.3 PCPMer – a method to identify sequence motifs based on PCPs 15

CHAPTER 2: ENGINEERING NOVEL PROTEINS WITH ENHANCED MECHANICAL STRENGTH **19**

2.1 Introduction 19

Elastic properties of proteins and Titin 19

Mechanical strength of Ig-like domains in Titin 21

Engineering protein domains with specified mechanical stability 23

2.2 Methods 25

Expression and purification of poly-proteins 25

CD spectroscopy 26

Atomic Force Microscopy 26

Homology modeling 26

Steered Molecular Dynamics simulation 27

2.3 Results 30

Design of Titin I1-like domains with enhanced mechanical stability 30

Stability of the engineered domains 33

The mechanical strengths of the engineered novel Titin domains 36

Steered Molecular Dynamics simulation of the engineered proteins 40

2.4 Discussion 44

2.5 Conclusions and future studies 47

CHAPTER 3: PREDICTION OF ALLERGENICITY 48

3.1 Introduction 49

Allergenic proteins 49

Assessment of allergenicity 51

3.2 Methods 52

SDAP 52

Datasets 53

Scoring method 54

Binding efficiency test 57

3.3 Results 59

Evaluation of FAO/WHO criteria 59

Allergenic motifs for 17 Pfam-A families 60

Motifs for protein families related to peanut allergens. 65

Scores of sequences from human microbiome 68

Correlation between the cross-reaction results of allergen chip and their allergenicity scores... 71

3.4 Discussion 75

3.5 Conclusions and future studies.. 79

CHAPTER 4: COMPUTATIONAL DESIGN OF VACCINES AGAINST MULTIPLE STRAINS OF ENCEPHALITIC ALPHAVIRUSES 80

4.1 Introduction 81

Alphaviruses 81

Encephalitic Alphaviruses, VEEV and EEEV 82

Envelope proteins of alphaviruses 86

4.2 Methods.. 90

Phylogenetic tree.. 90

PCP distance and similarity 91

InterProSurf.. 91

Construction of Hybrid Viruses 93

4.3 Results 94

Sequences of VEEV and EEEV E2 protein 94

Predicted epitopes on E2 protein of VEEV 98

Design of hybrid E2 protein.. 100

Homology models of Hybrid E2 proteins 105

4.4 Discussion 108

4.5 Conclusions and future studies.. 111

CHAPTER 5: SUMMARY and Future activities... 112

REFERENCES.. 114

List of Figures

Figure 1.1 workflow of PCPMer 18

Figure 2.1 Procedures of Steered Molecular Dynamics simulation 29

Figure 2.2 Sequences and Homology models of the four engineered proteins. 32

Figure 2.3 Minimized energies of the engineered proteins and wild types. 34

Figure 2.4 Far-UV CD spectra for the engineered proteins.. 35

Figure 2.5 Poly-protein and Force-extension profile of wild type I1. 38

Figure 2.6 Force-extension profiles of engineered proteins. 39

Figure 2.7 Force-extension profiles of SMD simulation 42

Figure 2.8 Pairwise interactions in SMD simulation 43

Figure 3.1 Peptide microarray assays 58

Figure 3.2 Evaluation of the FAO/WHO criteria of allergenicity. 60

Figure 3.3 Distribution of z-scores for non-allergenic sequences and allergenic sequences. 64

Figure 3.4 Cumulative percentage of non-allergenic sequences and allergenic sequences. 65

Figure 3.5 z-scores for allergenic sequences in PF00544 (Pectate Lyase) and their homologues in Human Microbiome. 71

Figure 3.6 2-dimensional MDS of distance matrix of cross-reactivity............................ 74

Figure 4.1 Overview of the strategy to design multivalent vaccine 81

Figure 4.2 Phylogenetic tree of major species of alphaviruses.. 86

Figure 4.3 Envelope glycoproteins of VEEV subtype II, Everglades. 89

Figure 4.4 Propensity scores of epitope and surface for 20 amino acids......................... 93

Figure 4.5 Plasmid map for expressing hybrid proteins 94

Figure 4.6 Sequence identity heat map of the E2 protein of VEEV. 96

Figure 4.7 Sequence identity heat map of the E2 protein of EEEV 97

Figure 4.8 Predicted antibody binding sites for E2 protein of VEEV and EEEV 99

Figure 4.9 Mapping conserved residues and predicted epitopes on VEEV and EEEV.. 101

Figure 4.10 The strategy to design hybrid E2 proteins.. 102

Figure 4.11 Experimental epitopes for VEEV and predicted epitopes for EEEV. 104

Figure 4.12 Sequences of designed hybrid E2 proteins. 106

Figure 4.13 Homology models of the hybrid E2 proteins. 107

List of Tables

Table 2.1 Residues involved in pairwise interactions.. 44

Table 3.1 Summary of allergens in top 17 Pfam-A families in SDAP 63

Table 3.2 Binding intensity of motifs from major peanut allergens to IgE 67

Table 3.3 Homologues of pectate lyase in Human Microbiome 70

Table 3.4 Scores of the 10 allergens in PF00407... 73

Table 3.5 Distance matrix generated from the cross-reaction results............................... 74

Table 3.6 Summary of SVM-based allergenicity prediction methods............................. 78

Table 4.1 Mutation sites in hybrid E2 proteins.. 106

Table 4.2 Assessment of the quality of the models 108

List of Abbreviations

AFM	Atomic Force Microscopy
CHIKV	Chikungunya virus
EEEV	Eastern Equine Encephalitis virus
Ig	Immunoglobulin
MSA	Multiple Sequence Alignment
PCP	Physical–Chemical Property
RMSD	Root Mean Square Deviation
SDAP	The Structural Database of Allergenic Proteins
SMD	Steered Molecular Dynamics
SVM	Support Vector Machine
VEEV	Venezuelan Equine Encephalitis virus
wt	Wild Type

PREFACE

DEDICATION

To my father Guojun, who has been working really hard to support the family for many years and taught me to become an independent man.

To my mother Yvming, who has been taking great care of the whole family and built us such a sweet home.

To my little sister Wenhui, who has been a great companion and brought so much laughter through our childhood.

To my grandpa Zhipeng, who is a wise man and shared lots of his life experiences with me.

THANKS

The road to a PhD is never easy. Graduate school is full of fun and also challenges. A lot of faculty, staff and friends have generously helped me through this tough adventure. First of all, I would like to sincerely thank my mentor Dr. Werner Braun, who is always patient with me and willing to discuss my projects. He has maintained a great research environment and offered many inspiring advices at various aspects of my research.

Secondly, I want to thank all the members of my committee. They did great job at monitoring my research progress and giving suggestions. Dr. Andres Oberhauser is an expert in Atomic Force Microscope and we had a great collaboration in the project to engineer novel Ig domains in Titin. As the chair of my committee, Dr. Junji Iwahara helped organize the committee meetings and led the discussion. Dr. Malgorzata Rowicka discussed with me about the scoring methods in the project of predicting allergenicity. Dr. Montgomery Pettitt has extensive experiences in computational chemistry and he offered many great advices in molecular dynamics simulations. Dr. Randall M. Goldblum has worked on allergy for many years and shared his knowledge on the project of predicting allergenicity.

Thirdly, I want to thank all the collaborators, postdocs, graduate students and technicians. They have been very helpful through my projects. Dr. Catherine Schein and Dr. Surendra Negi worked with me on multiple projects and publications. Dr. Liang Ma from Dr. Oberhauser's group helped me with the molecular biology part during engineering novel Ig domains. Dr. Naomi Forrester co-mentored me on the project to computationally design vaccines against alphaviruses. The staff and administrators in graduate school, BMB and MBET program have made a lot of things much easier to handle. I would like to thank Debora Botting, JoAlice Whitehurst and Lisa Pipper who are always nice and willing to help out. I highly appreciate the funding support from the IHII / SCSBMB Joint predoctoral fellowship.

Finally, I would like thank my girlfriend and family for always being there for me and supporting me.

ABSTRACT

The functions and properties of proteins are mainly determined by their amino acid sequences and 3-dimensional structures. Motifs composed of amino acids that are either continuous in sequence or cluster in space play important roles in structures and functions of proteins. The functional annotation of these motifs provides insights into the mechanism of protein structures and functions and therefore contributes to our understanding of the protein folding process, and aids in the rational design of new therapeutic drugs and vaccines. In my research, I finished three projects where I identified specific motifs based on the conservation of Physical–Chemical Properties (PCP) of amino acids that could be related to the mechanical strength of proteins, to the potential allergenicity of proteins and to the antigenic diversity of the envelope proteins of encephalitic alphaviruses. I engineered novel Immunoglobulin (Ig) domains of the muscle protein titin by introducing motifs from mechanically strong Ig domains to weak Ig domains. The Atomic Force Microscope (AFM) results showed the enhanced mechanical strength for several of the engineered proteins. I developed a scoring method to evaluate the allergenicity of query sequences based on the characteristic motifs that are specific for allergenic sequences. The validation on peanut allergens, the family of pectate lyase showed its ability to distinguish allergenic sequences from non-allergenic ones. In addition, I identified conserved antibody binding epitopes on the envelope E2 proteins of diverse encephalitic alphaviruses and designed hybrid vaccines against multiple strains of alphaviruses. Through these projects, I demonstrated that mining PCP motifs is a practical computational approach to deepen our understanding of protein

functions, to generate specific hypothesis for experimental verification and to guide the rationale design of novel vaccines.

CHAPTER 1: INTRODUCTION

1.1 Protein sequence motifs

Proteins are fundamental macromolecules involved in virtually every biological process in living organisms. The sequential arrangement of amino acids provides the essential information for the structures and functions of proteins. While an enormous amount of different proteins are discovered and the number is still increasing due to the advance in sequencing technology, proteins with similar structures or functions usually share similarity in their sequences. Many databases have been established to organize different proteins by protein families according to their homology in sequences, functions or structures, such as PROSITE (Sigrist et al., 2013), Pfam (Finn et al., 2014) or SCOP (Murzin et al., 1995). Most of these databases also contain sequence profiles or hidden Markov models generated from multiple sequence alignments to represent the common features of protein families and enable search of homologous proteins (de Castro et al., 2006; Finn et al., 2014; Gough et al., 2001).

In addition to overall similarity of sequences, homologous proteins also contain critical residues that are either continuous in sequence or close in space playing important roles in maintaining stable folds or conducting functions for proteins. Due to their importance in structure or function, these residues are usually conserved over large taxonomic distances and evolutionary time among homologous proteins (Brutlag, 2006). These residues are often referred to as sequential or conformational motifs. The definition of protein sequence motifs is rather vague and it mainly implies the conservation of amino acids. Some motifs are rather universal and very specific for certain functions such as glycosylation sites; meanwhile, most motifs represent the characteristic features of a specific set of homologous proteins (Bork and Koonin, 1996). Identification of these motifs would significantly deepen our understanding of the structures and functions of proteins. Many bioinformatics tools have been developed to generate characteristic motifs for homologous proteins and search for the existence of motifs defined by either experimental results or computational methods (Attwood et al., 2012; Bailey et al., 2009; de Castro et al., 2006; Dinkel et al., 2014; Edwards et al., 2007; Mathura et al., 2003; Mi et al., 2012). For example, MEME is a motif discovery method based on probabilistic algorithms (Bailey et al., 2009; Bailey and Elkan, 1994). The method breaks given set of sequences into segments of certain length and infers the probabilities of each segment being in a motif or background by iterations of expectation-maximization algorithm. However, motifs generated by MEME are usually long segments and the efficiency of the program is decreased with smaller window sizes (Bailey and Elkan, 1994).

1.2 Physical–Chemical Properties (PCP) of amino acids

Amino acids are the basic building blocks for all proteins that are responsible for proper biological functions. Each one of the 20 standard amino acids has unique physical and chemical properties (PCP) that determine its role in the folding and function of proteins. For example, hydrophobic residues are more likely to be located in the core of protein to stabilize the structure while charged and polar residues are often found exposed on the surface or involved in the interfaces of protein-protein interaction (Negi and Braun, 2007; Tsai et al., 1997). However, it becomes extremely complicated to analyze when the length of the amino acid polymer extends and the number of physical-chemical properties describing each amino acid increases (Mathura and Braun, 2001). While PCPs are composed of measurements in many different aspects, there is a significant amount of redundant information in PCPs, such as molecular weight and length of side chain which are positively correlated with each other. Therefore, the analysis of protein sequences can be simplified by reducing the redundancy in PCPs while still summarizing the properties of amino acids precisely.

In a previous work, Braun et.al developed a new method to summarize the PCPs of 20 naturally occurring amino acids (Mathura and Braun, 2001). They used the method of multidimensional scaling on the original set of 237 PCPs of each amino acid, and constructed a 5 dimensional property space in which the spatial distribution of amino acids is highly similar to the distribution in original high-dimensional property space. Finally, 5 quantitative descriptors, E1 through E5, were generated to represent the PCPs for each amino acid. The physical meaning of each descriptor can be correlated with a certain type of property; however, they cannot simply be replaced by corresponding individual properties. E1 correlates with the hydrophobic/hydrophilic property and E2 with the size of the amino acids. The E3 vector correlates with several helical propensity scales. E4 is partially related to the partial specific volumes, number of codons and relative abundance of amino acids. E5 correlates weakly with beta-strand propensity. The 5 quantitative descriptors significantly simplify the trouble of dealing with high-dimensional property space while keeping the essential information.

1.3 PCPMer – a method to identify sequence motifs based on PCPs

Homologous proteins with similar function or properties usually share similar sequences. Even when the overall sequence similarity is as low as 20 -30 %, some of the critical regions are still conserved to maintain similar folds and functions (Abagyan and Batalov, 1997). These conserved regions are of great importance for the structure and function of proteins and they can provide insights into the studies of protein folding, functional sites discovery or drug/vaccine development. Various methods have been developed to identify these regions and many of them have achieved great success. Conventional methods to identify motifs usually only focus on the composition or frequency of conserved amino acids (Attwood et al., 2012; Bailey et al., 2009; de Castro et al., 2006; Dinkel et al., 2014; Edwards et al., 2007; Mi et al., 2012), while in reality, a more reasonable assumption is that, as long as the physical–chemical properties of the critical residues are still conserved, the amino acid type is allowed to change during evolution. Therefore, novel method needs to be developed to identify rather subtly motifs which are conserved in properties.

The 5 quantitative descriptors generated by Braun laboratory, E1 through E5, summarize the physical-chemical properties of 20 standard amino acids (Mathura and Braun, 2001). These descriptors have provided us with a set of simplified yet precise parameters to compare the similarity of amino acids in terms of their PCPs. The software package PCPMer was developed to generate physical-chemical conserved sequence motifs and search for matches of these motifs in query sequences (Mathura et al., 2003). A web interface was made publicly available at http://landau.utmb.edu:8080/pcpmer/index.jsp.

The workflow of the software is shown in Figure 1.1. The software takes multiple sequence alignment of protein families or related proteins as input and produces a set of physical-chemical conserved sequence motifs and their quality scores as output. The criteria for conservation are derived from the distributions of the PCP descriptors in each column of the alignment as compared to the background distributions. The range of each descriptor is divided into five equal bins, and the distributions are calculated according to these bins. The background distributions are calculated using non-redundant proteins from the Swiss-Prot database (Bairoch et al., 2004) as a random sample. To compare the distribution of E1-5 descriptors as found in a given column *j* of the multiple sequence alignment to the background distribution, the relative entropy (or Kullback-Leibler divergence) (Kullback and Leibler, 1951) for each of the five vectors E1 to E5 is calculated by

$$\boldsymbol{K_n^j = \sum_{i=1}^{5} Q_n(i) log_2 \frac{Q_n(i)}{P_n(i)}}$$

Where *j* is the index of the column number in multiple sequence alignment, the index *n* iterates over the five descriptors E1 to E5 and the index *i* over the five bins. $Q_n(i)$ gives the percentage of the *n*-th E value observed in the bin *i* for the alignment and $P_n(i)$ is the corresponding percentage in the background distribution. The final relative entropy score of column *j* of the alignment is calculated by summing the scores for five descriptors.

$$\boldsymbol{K^j = \sum_{n=1}^{5} K_n^j}$$

The relative entropy score measures the difference between the observed distribution in multiple sequence alignment and the background distribution. For the alignment, if the

PCPs of the amino acids in a column are highly conserved, the distributions of the descriptors E1-5 are significantly non-uniform comparing with the background distribution, resulting in a high relative entropy score. If the distributions of descriptors in alignment are the same as the background distribution, the relative entropy score is 0. The motifs are defined as continuous stretches of residues with relative entropy values higher than an empirical or user specified threshold. The motifs are typically 5-15 columns in length with minimum length and inclusion of gaps decided by the user. Rather than an alphabetical peptide, each motif is represented by quantitative profiles including the average values, standard deviations and the relative entropy scores for each descriptor E1-5 at each position of the motif.

While in most studies the motifs are defined by the absolute conservation of particular amino acid types, the PCP approach allows a quantitative precise definition of sequence motifs that also can be applied to subtle functional important motifs. PCPMer has been successfully applied in various studies (Garcia et al., 2009; Ivanciuc et al., 2009a; Ivanciuc et al., 2004; Lu et al., 2012; Mathura et al., 2003; Oezguen et al., 2008a; Schein et al., 2005a; Schein et al., 2002; Schein et al., 2005b). For example, the work on Apurinic/apyrimidinic Endonuclease (APE) protein family by Braun laboratory has shown the ability of PCPMer to identify meaningful PCP motifs (Mathura et al., 2003). They collected the sequences of homologous APE from a wide range of organisms from prokaryotes to eukaryotes and generated 12 PCP motifs with PCPMer. They found out that all residues known to be involved in metal ion binding of APE are included in the PCP motifs and three other motifs defined by PROSITE also match with their PCP motifs.

In my studies, I extended the application of PCPMer to generating characteristic PCP motifs that are involved in the mechanical strength, allergenicity and antibody binding epitopes of proteins. In all three applications, the use of the PCP motifs goes beyond sequence similarity search, by directing new experimental studies and designing novel proteins with improved functional properties. These successful applications demonstrated that PCP motifs mining is a powerful general method to expand our understanding of the mechanisms of protein folding and functions.

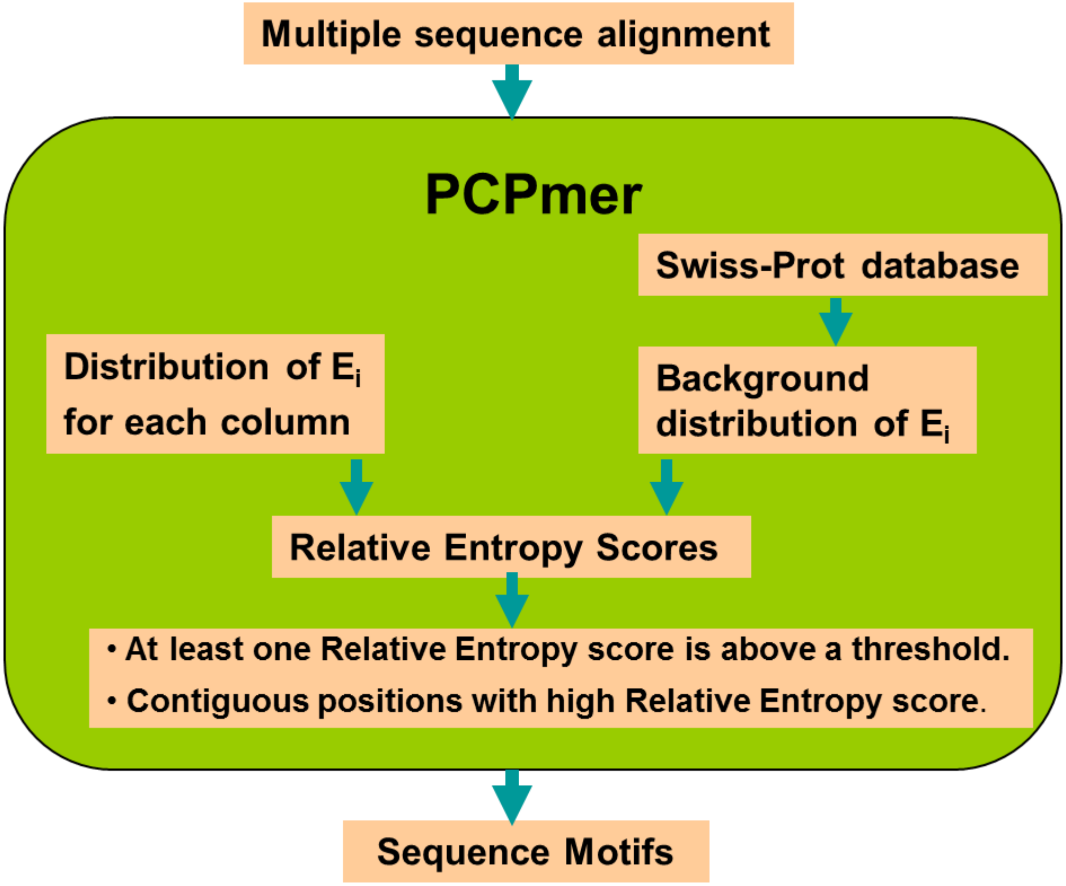

Multiple sequence alignment
PCPmer
Swiss-Prot database
Distribution of E_i for each column
Background distribution of E_i
Relative Entropy Scores
• At least one Relative Entropy score is above a threshold.
• Contiguous positions with high Relative Entropy score.
Sequence Motifs

Figure 1.1 workflow of PCPMer

CHAPTER 2: ENGINEERING NOVEL PROTEINS WITH ENHANCED MECHANICAL STRENGTH

The use of atomic force microscopy (AFM) has recently led to a better understanding of the molecular mechanisms of the unfolding process by mechanical forces; however, the rational design of novel proteins with specified mechanical strength remains challenging. We have approached this problem from a new perspective by generating linear physical-chemical properties (PCP) motifs from sequences with known AFM data. Guided by our linear sequence analysis, I designed and analyzed four new mutants of the Titin I1 domain with the goal of increasing the domain's mechanical strength. All four mutants could be cloned and expressed as soluble proteins. AFM results indicated that at least two of the mutants have increased molecular mechanical strength. This observation suggests that the PCP method is useful to graft sequence fragments specific for high mechanical stability to weak proteins to increase their mechanical stability, and represents an additional tool in the design of novel proteins besides steered molecular dynamics calculations, coarse grained simulations and phi-value analysis of the transition state.

2.1 Introduction

Elastic properties of proteins and Titin

Many cellular activities of proteins, such as such cell adhesion, translocation or ligand binding, require specific mechanical stabilities of the protein domains (Bustamante et al., 2004; Forman and Clarke, 2007; Galera-Prat et al., 2010; Oberhauser and Carrion-

Vazquez, 2008; Puchner and Gaub, 2009). Interestingly, some proteins with no obvious mechanical functions, like the B1 domain of protein G, also show a remarkably high mechanical stability (Li et al., 2006). Current experimental data on the mechanical strength of about 100 protein domains is collected in the BSDB database (Sikora et al., 2011). There has been great progress in understanding how proteins react to external mechanical forces, using single molecule force spectroscopy by atomic force microscopy (AFM) or optical tweezers, combined with steered molecular dynamics (SMD) calculations in recent years (Galera-Prat et al., 2010; Hsin et al., 2011), yet general rules that govern the mechanical stability of proteins are still elusive.

The giant muscle protein Titin is one of the extensively studied proteins with unique mechanical properties. Also known as connectin, Titin is the longest covalently linked protein known in the human genome with ~30,000 amino acids (Lander et al., 2001; Wang, 1996). A single Titin molecule spans half of the sarcomere through I-band and A-band. The ends of Titin are connected to the Z disc and M line of sarcomere. The main function of Titin molecule is to exert a passive force that keeps sarcomere components uniformly organized during muscle stretching and restores sarcomere length when the muscle is relaxed (Improta et al., 1996; Lu et al., 1998; Marszalek et al., 1999; Mayans et al., 2001; Pfuhl et al., 1997; Williams et al., 2003).

The spring-like molecule Titin is primarily composed of ~300 tandem repeats of two types of domains, Immunoglobulin-like (Ig-like) and fibronectin type III (Fn3) domains. The I-band region of Titin is thought to be responsible for its passive elasticity. In cardiac muscle, the Titin I-band contains four structural units: the proximal Ig-like domains, the N2B or N2BA segment, the PEVK region, and the distal Ig-like domains. 70% of the

PEVK region is composed of proline, glutamate, valine, and lysine. It forms a mixture of unstable coiled conformations and polyproline type II helix and easily elongates to increase the length of sarcomere under small forces (Granzier and Labeit, 2004; Linke, 2008; Wang, 1996). The structure of the N2B and N2BA segments is still unknown. There are many binding sites for signaling molecules in these segments, suggesting that they may be involved in signaling pathways (Granzier and Labeit, 2004; Kruger and Linke, 2009; Kruger and Linke, 2011; Linke and Hamdani, 2014; Linke and Kruger, 2010). The proximal and distal Ig-like domains are the main regions that provide elastic properties to Titin.

Titin has been associated with several diseases, such as dilated cardiomyopathy (DCM) (Seidman and Seidman, 2001), hypertrophic cardiomyopathy (HCM) (Satoh et al., 1999) and tibial muscular dystrophy (TMD) (Hackman et al., 2002). DCM and associated heart failure are major causes of human morbidity and mortality. The reported disease-causing missense mutations are found in Ig domains located at the Z-disk region, M-line region and the elastic I-band of Titin (Kruger and Linke, 2009; Linke and Hamdani, 2014). Mutations may uncouple domains from their binding partners and lead to stress-sensing defects, myocyte dysfunction and cardiac chamber dilation (Kruger and Linke, 2011).

Mechanical strength of Ig-like domains in Titin

The Ig-like domains in Titin are important contributors to its elastic property. There are about 40 Ig-like domains in the I-band of Titin (I1, I2 … I40), and each has ~100 amino acids. They share a common beta-sandwich fold with two beta-sheets packing against each other. The structure is stabilized by hydrophobic interactions between the two beta-sheets and by inter-strand hydrogen bonds (Mayans et al., 2001; Stacklies et al., 2009).

Single-molecule AFM is a powerful, high-resolution tool for imaging and force measurement of macromolecules. In AFM experiments, a single poly-protein composed of tandem repeat domains is fixed to the substrate at one terminal and extended by external forces at the other terminal at constant velocity. The domains unfold successively, resulting in a force-extension profile where the forces needed to unfold them are recorded as a saw-tooth shape curve along the extension length and each of the peaks corresponds to the rupture of one domain. The measured forces are positively correlated with the pulling velocity, and it was suggested to be proportional to the square root of the velocity (Hummer and Szabo, 2003). Previous AFM data on measuring the unfolding forces of Ig domains showed a mechanical hierarchy in the unfolding forces of Ig-like domains, where domains in the distal region unfold at forces two to three times greater than domains in the proximal region of the I-band (Garcia et al., 2009; Li et al., 2000; Li and Fernandez, 2003; Li et al., 2002). For example, the proximal domain I1 unfolds at force of ~100pN while the distal domain I27 unfolds at ~200pN at pulling velocity of ~0.5nm/ms in AFM experiments (Li et al., 2000; Li and Fernandez, 2003).

In addition to AFM experiments, SMD simulations were used to analyze the details of the unfolding process of Ig-like and fibronectin domains in Titin. The simulations on I1 and I27 provided evidence that the topology of secondary structures plays an important role in the mechanical strength (Gao et al., 2002a; Gao et al., 2002b; Lu et al., 1998; Paci and Karplus, 1999). It was observed that the β-strands A-B and A’-G located near the N and C terminus form two sets of hydrogen bonds, which function as a mechanical clamp that resists the shear forces applied to the termini. The importance of the hydrogen bond network was verified by AFM experiments on mutant proteins with point mutations that disturb these hydrogen bonds (Li et al., 2000; Li and Fernandez, 2003). This type of shearing motif or mechanical clamp is also found in other β-sandwich folds like fibronectin III domains (Paci and Karplus, 1999) and some other mechanically stable folds like the ubiquitin-like (β-grasp families) protein GB1 (Li et al., 2006) and microbial cellulosome protein scaffoldins (Valbuena et al., 2009). Thus it was proposed that the shear mechanical clamp was responsible for the main force barrier observed for Ig domains in the AFM experiments.

Engineering protein domains with specified mechanical stability

The rational design of proteins with specified molecular strengths needs a comprehensive understanding of the contribution of individual residues to the experimentally observed mechanical strength. A particular challenge is to design novel proteins with increased mechanical strength rather than to reduce strengths by disturbing the original structure (Borgia et al., 2008; Crampton and Brockwell, 2010; Galera-Prat et al., 2010).

Based on the previous knowledge of the mechanism of mechanical stability, several studies have been conducted on various types of domains to manipulate their mechanical strengths. Clarke's group used the analysis of hydrophobic core packing as a guide in the re-engineering of a fibronectin type III domain (Ng et al., 2007). The side chain core of the third domain (TNfn3) of fibronectin was compared to that of the homologous tenth domain (FNfn10). The hydrophobic core of FNfn10 was then replaced with that of the homologous, mechanically stronger TNfn3 domain, increasing the mechanical stability of the engineered FNfn10. A recent study of the macro domain AF1521, a protein from the thermophilic bacteria *Archaeoglobus fulgidus*, showed that water-accessibility of the load bearing hydrogen bonds correlates with mechanical stability (Guzman et al., 2010).

Another interesting attempt was by Li group to engineer metal chelation sites into various locations of protein GB1. The binding of metal ions stabilized native state over unfolding transition state (Cao et al., 2008). Another work on bacterial protein L with β-grasp fold (ubiquitin-like) by Sadler et al. found that the size of the hydrophobic side-chain at position 60 plays a critical role in its mechanical strength. Mutation of I60V resulted in a decrease of ~36pN, while I60F resulted in a ~72pN increase in mechanical strength (Sadler et al., 2009)

In addition to the previously reported mechanical clamp for the Ig domains in Titin, multiple studies have shown evidence of interactions from other regions contributing to the mechanical strengths. Li et al. (Li et al., 2000) did proline mutagenesis in the force-bearing regions of I27 aiming at breaking the hydrogen bonds and reducing its strength; however, one of the mutations at Y9P in strand A' accidently increased the strength of I27. Sharma et al. (Sharma et al., 2006) shuffled the A'-G strands from stronger domain I32 (~300pN) to I27 (~200pN) in an attempt to generate hybrid domains with mechanical stabilities higher than I27. Their unsuccessful design indicates that the mechanical clamp is not the only structural region responsible for the mechanical stability of Titin Ig domains. Another study involving more residues which are close to the A'-G strands is successful, suggesting the side-chain interactions with the neighboring residues are also important (Borgia et al., 2008).

All these examples suggest that we are far from a comprehensive understanding of the molecular nature of mechanical stability, and there are no established approaches to engineer homologous domains with a specified mechanical strength.

Our design strategy used force-specific linear sequence motifs that were recently determined for the Titin Ig domains (Garcia et al., 2009). The approach took advantage of the differences in the mechanical stabilities of the Ig domains in the I-band of Titin, where the proximal Ig domains close to the N-terminus (weak domains) unfold at forces much lower than the distal Ig domains close to the C-terminus of the I-band (strong domains) (Li et al., 2002). Specific PCP motifs were identified for strong and weak Ig domains respectively by analyzing homologous sequences across species using PCPMer (Mathura et al., 2003). Four recombinant proteins were designed by introducing strong motifs or structural fragments from strong Ig domain I27 to weak domain I1. The validation by AFM experiments confirmed that at least two of the recombinant proteins have increased mechanical strength. This encouraging result represents the first step to further delineate the contribution of individual motifs to the overall molecular strength and to establish a quantitative predictive method of molecular strength for domains with identical folds but large sequence variations.

2.2 Methods

Expression and purification of poly-proteins

The DNA sequences for the four engineered mutants of the I1 domain were synthesized by DNA2.0 Inc (Menlo Park, CA). Poly-proteins with five sequential arranged domains were constructed for each designed domain to measure the effects on the mechanical stability by AFM experiments. Each poly-protein is composed of four wild type I27 domains flanking one designed target domain in the middle. The wild type I27 domain is used as an internal fingerprint since its mechanical property is well characterized. A 6-

His-tag was attached to the N-terminal for purification and anchoring to the cover slip. The dipeptide Cys-Cys was added to the N-terminal for the attachment to the AFM probe. The poly-proteins were expressed in *E.coli BL21* using home-made plasmid *pAFM* (Steward et al., 2002) and purified by Ni-affinity chromatography.

CD spectroscopy

Circular dichroism (CD) spectroscopy of all the poly-proteins were recorded with *JASCO J-815* (JASCO Inc., Easton, MD) at room temperature. A 10mM Phosphate Buffer was used as buffer for the proteins and its signal was later subtracted from the spectra. The path length of the cuvette was 1mm. The concentrations of proteins were determined by Bradford assay for the fitting of the CD spectra. Each spectrum was averaged from three scans from 200nm to 260nm at scanning speed of 50nm/min.

Atomic Force Microscopy

The unfolding forces of the novel proteins were measured with our home-built AFM instrument at room temperature of 25 Celsius degrees. The poly-proteins were diluted with 10mM PBS buffer, and 10mM Dithiothreitol (DTT) was added to the solution to prevent poly-proteins from forming dimmers. A drop of ~10μl solution was incubated on the Ni-coated cover slips for 15min and then rinsed with PBS to remove unbound poly-proteins. The AFM probe used in the experiments was MLCT (Bruker, CA) with typical spring constant of 0.03N/m, and the pulling speed was ~0.5nm/ms.

Homology modeling

Three-dimensional structural models for 38 Ig domains with unknown 3D structures in the I-band of cardiac Titin and the four engineered domains were generated with standard methods successfully used in our laboratory in the past (Ivanciuc et al., 2004; Oezguen et al., 2008b). First, the sequences of the 38 Titin domains were submitted to the meta-server, Genesilico (Bujnicki et al., 2001) to search for the best template structure. All possible templates obtained from the meta-server were analyzed manually and the best template was selected for homology modeling. Homology model structures were generated using our homology modeling software package MPACK, and the software package modeller (Eswar et al., 2006). The model structures obtained from MPACK were further energy minimized using the FANTOM program. The root mean square deviation (RMSD) value between the template and model structures obtained from MPACK and modeller was found to be less than 0.5 Å. Finally, the model structures and the X-ray crystal structures of I1 and I27 were energy minimized in a water box with TIP3P water molecules for 10000 iterations with the NAMD software (Phillips et al., 2005) to test the stability of the model structures. The RMSD between the final model structures and the templates was calculated by Pymol (The PyMOL Molecular Graphics System, Version 1.3, Schrödinger, LLC).

Steered Molecular Dynamics simulation

The Steered Molecular Dynamics (SMD) simulation of the unfolding process of wild type and engineered Ig domains under external forces were finished using the package NAMD (Phillips et al., 2005; Wang et al., 2011) with the CHARMM22 force field. Unlike the AFM experiment where the unfolding of a five-domain poly-protein is measured, SMD only simulates the unfolding of one domain at a time due to the limit of computing power. Starting from the PDB coordinates file of the one Ig domain, firstly, the hydrogen atoms were regenerated by VMD (Humphrey et al., 1996) to enable the evaluation of the impact of hydrogen bonds during unfolding. The Ig domain was then placed in the center of a water sphere with a radius of 40Å (Figure 2.1 A). The temperature of the whole system was increased to 300K to mimic the room temperature in AFM experiments. Then the energy of the system was minimized for 1000 steps and equilibrated for 10ps at the time step of 1 fs/step. After a stabilized all atom system was reached, the coordinates of the C-alpha of the N terminus was fixed to prevent it from any movement, and an external force was applied to the C-alpha of the C terminus in the direction of N terminus to C terminus (Figure 2.1 B). The external force is simultaneously adjusting to keep the C terminus of Ig domain moving at a constant velocity, thus simulating the unfolding process as in AFM experiments. I recorded the unfolding trajectory of each Ig domain and generated force-extension profiles to analysis the key events during unfolding. Multiple simulations were finished on I1 wild type, I27 wild type and engineered protein I1s1s2 at the constant velocities of 0.5Å/ps, 0.1Å/ps and 0.01Å/ps. The forced unfolding was simulated to reach the extension of at least 25Å to ensure the rupture of the beta sandwich fold. In addition to the analysis of overall force-extension profiles, the interaction energies between each pair of residues that are involved in previously reported force-bearing regions were monitored under the velocity of 0.01Å/ps.

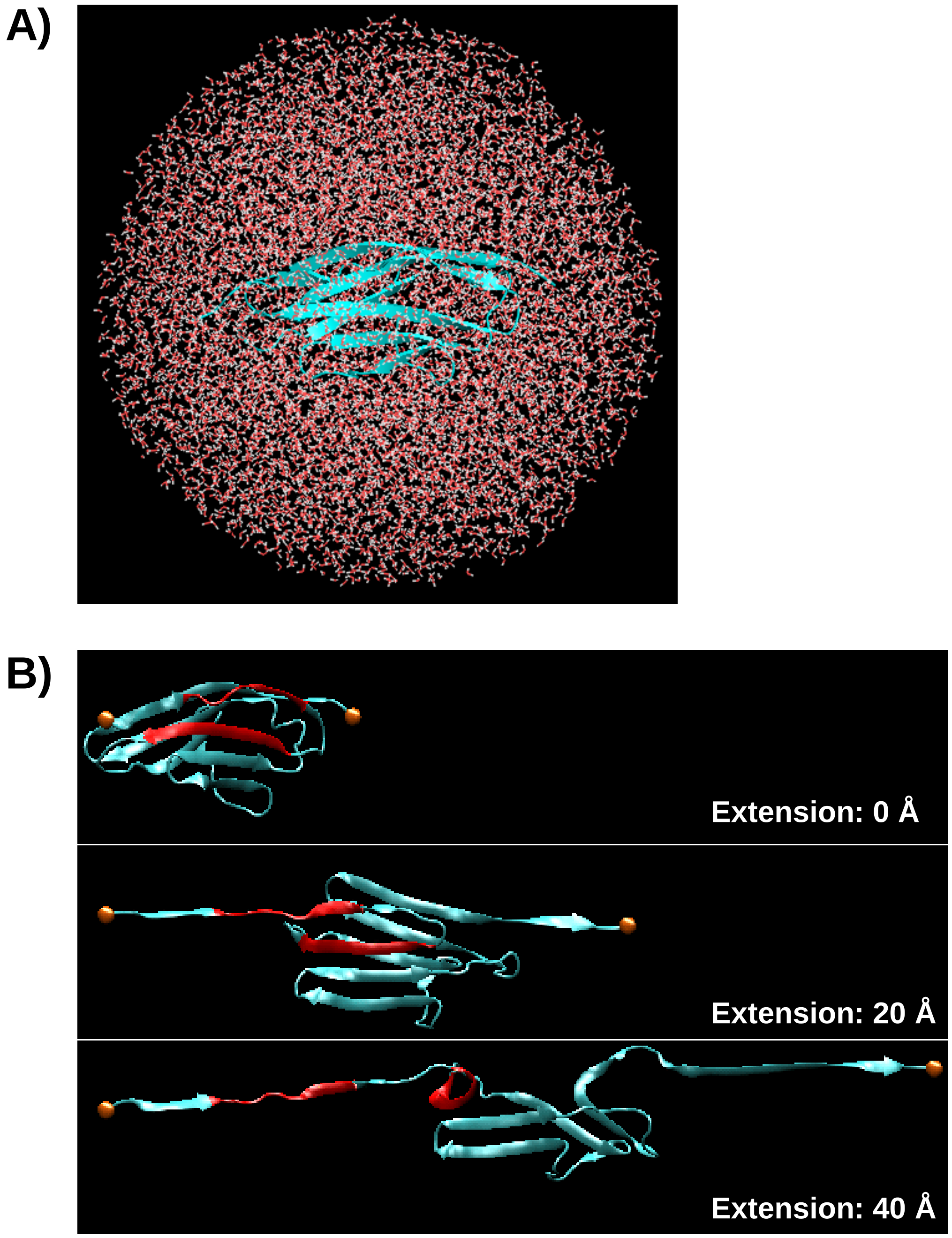


Figure 2.1 Procedures of Steered Molecular Dynamics simulation

The SMD simulation of the unfolding process of I27 is used for demonstration.

A). Ig domains are solved in water sphere with a radius of 40Å

B). N terminus is fixed while C terminus is pulled at constant velocity. The previously suggested force-bearing regions are highlighted in red.

2.3 Results

Design of Titin I1-like domains with enhanced mechanical stability

The sequences of different Ig domains in the I-band of cardiac muscle protein Titin are quite diverse with an average sequence identity of only 25% (Garcia et al., 2009). On the other hand, the available experimental 3D structures of Titin Ig domains, such as the NMR and X-ray structure of I27 (Improta et al., 1996; Stacklies et al., 2009) and, the X-ray crystal structure of I1 (Mayans et al., 2001) and our comprehensive 3D modeling for all other 38 Ig domains in cardiac muscle indicate that all Titin Ig domains share the same Ig-like fold despite the large variation of amino acid sequences. A previous study by Garcia et al. analyzed the homologous sequences of Ig domains in Titin from up to 10 different species (Garcia et al., 2009). They grouped the sequences into weak domains I1, I4, I5 and strong domains I27, I28, I32, I34, and generated multiple sequence alignments for each group of homologous sequences across different species. The conserved PCP motifs for weak and strong Ig domains were identified by the software PCPMer (Mathura et al., 2003). While some of the motifs are common to both groups, suggesting their roles in maintaining the Ig-like fold, motifs that are unique for each group were recognized. Those motifs are considered as specific sequence signatures characterizing the mechanical stability of weak and strong Ig domains.

These unique motifs, two strong motifs s1 and s2, and two weak motifs, w1 and w3 are mapped on the amino acid sequences of wild type I27 and I1 domains (Figure 2.2 A). The aim of the design is to increase the mechanical strength of the weak domain I1 by introducing specific motifs or segments encompassing these motifs from the strong domains I27. Previous AFM studies have shown that the force needed to unfold wild type I1 is ~100pN at pulling velocity of ~0.5nm/ms (Li and Fernandez, 2003), which is significantly smaller than the unfolding force of ~200pN for I27 at same velocity (Li et al., 2000). A total number of four recombinant proteins were designed based on weak domain I1. The experimental structures of I1 (PDB: 1G1C) and I27 (PDB: 1WAA) were superposed by DaliLite (Holm and Park, 2000) to identify the residues that correspond to each other in structure. In the first recombinant I1s1s2, homologous positions in I1 were substituted by corresponding residues in the strong motifs s1 and s2 of I27. To subtract the potential impacts of the weak motifs, the second construct I1s1s2-w1w3 made additional changes based on I1s1s2 by replacing the weak motifs w1 and w3 in I1 with corresponding residues in I27. In the third recombinant I1_I27AB, the segment of A-A'-B strands which overlaps with the strong motifs and was previously reported as the force-bearing regions was introduced from I27 to I1. Finally, to ensure the hydrophobic interactions in the core are also considered, I1_I27AD was constructed by combining A-D strands of I27 with I1, where strand D overlaps with weak motif w3 and strand B, C are involved in the hydrophobic packing.

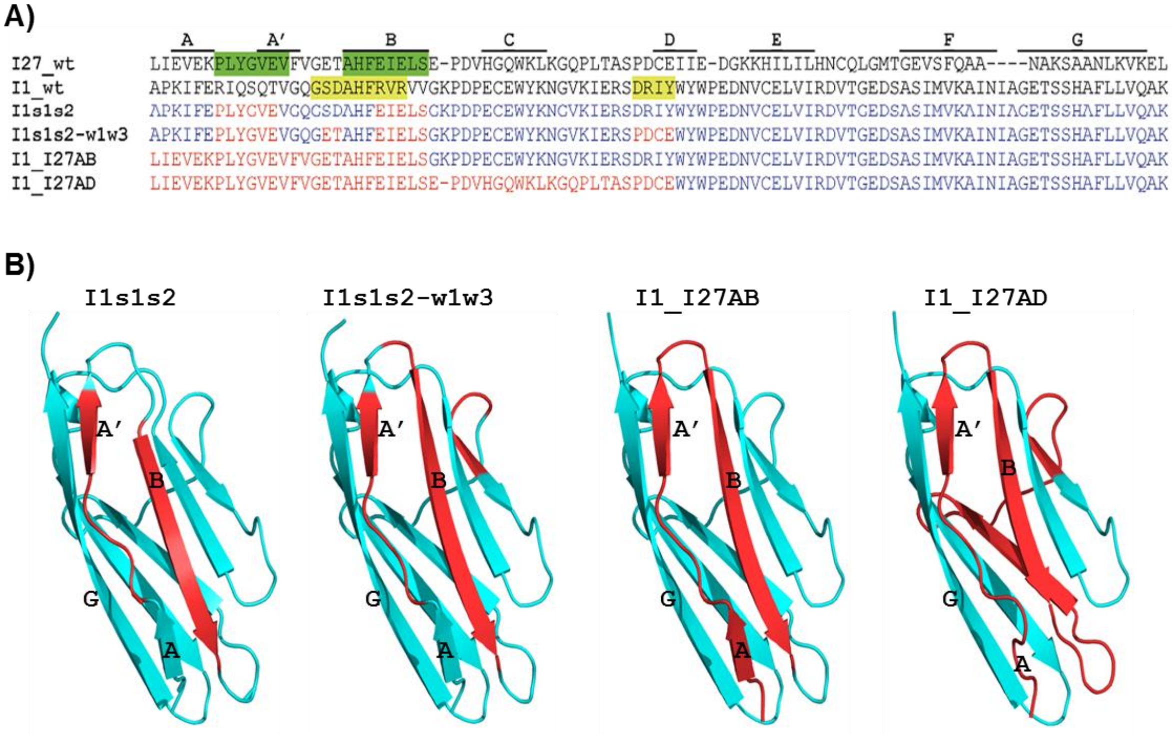

A)
A
A′
B
C
D
E
F
G
I27_wt LIEVEKPLYGVEVFVGETAHFEIELSE-PDVHGQWKLKGQPLTASPDCEIIE-DGKKHILILHNCQLGMTGEVSFQAA----NAKSAANLKVKEL
I1_wt APKIFERIQSQTVGQGSDAHFRVRVVGKPDPECEWYKNGVKIERSDRIYWYWPEDNVCELVIRDVTGEDSASIMVKAINIAGETSSHAFLLVQAK
I1s1s2 APKIFEPLYGVEVGQGSDAHFEIELSGKPDPECEWYKNGVKIERSDRIYWYWPEDNVCELVIRDVTGEDSASIMVKAINIAGETSSHAFLLVQAK
I1s1s2-w1w3 APKIFEPLYGVEVGQGETAHFEIELSGKPDPECEWYKNGVKIERSPDCEWYWPEDNVCELVIRDVTGEDSASIMVKAINIAGETSSHAFLLVQAK
I1_I27AB LIEVEKPLYGVEVFVGETAHFEIELSGKPDPECEWYKNGVKIERSDRIYWYWPEDNVCELVIRDVTGEDSASIMVKAINIAGETSSHAFLLVQAK
I1_I27AD LIEVEKPLYGVEVFVGETAHFEIELSE-PDVHGQWKLKGQPLTASPDCEWYWPEDNVCELVIRDVTGEDSASIMVKAINIAGETSSHAFLLVQAK
B)
I1s1s2
I1s1s2-w1w3
I1_I27AB
I1_I27AD
A′
B
G
A
A′
B
G
A
A′
B
G
A
A′
B
G
A

Figure 2.2 Sequences and Homology models of the four engineered proteins.

A). Two unique motifs of the strong Ig domains are mapped on the sequence of wild type I27 (s1, s2, highlighted in green). Two unique motifs of the weak Ig domains are mapped on wild type I1 (w1, w3, yellow). The sequences of the four engineered proteins are listed below where blue residues are from I1 and red residues are from I27.

B). Homology models for I1s1s2, I1s1s2-w1w3, I1_I27AB and I1_I27AD. The regions substituted by residues from I27 are shown in red.

Adapted from (Lu et al., 2012).

Stability of the engineered domains

To assess the stability of the designed Ig domains, homology models for all four engineered proteins were built using I1 (PDB: 1G1C) as template (Figure 2.3 B). All recombinant proteins properly fold into β-sandwich structures with small RMSD values comparing with the template (0.61, 0.68, 0.59 and 0.88Å). In addition, energy minimization was finished by NAMD for the engineered proteins, and also for wild type I1, I27 and models of the other 38 Ig-like domains in Titin I-band to provide a context for the final energies. The scale of energies indicates that all four designed domains have similar low energy values as the other wild type Ig domains of cardiac Titin, suggesting that the designed domains form stable folds (Figure 2.3).

In addition to simulations, the engineered proteins were expressed and purified as poly-proteins containing 5 Ig domains. CD spectroscopy was used to experimentally determine the folding of the recombinants. I measured far-UV CD spectra to evaluate the extent of the secondary structures using the poly-protein containing wild type I1 as control. All the engineered proteins showed spectra of similar patterns that are characteristic for β-strands (Figure 2.4), indicating their topology is mainly composed of β-strands. The spectra are also similar between the engineered proteins and wild type I1, suggesting that the designed proteins have similar β-sandwich folds to I1 and recombination doesn't cause noticeable misfolding.

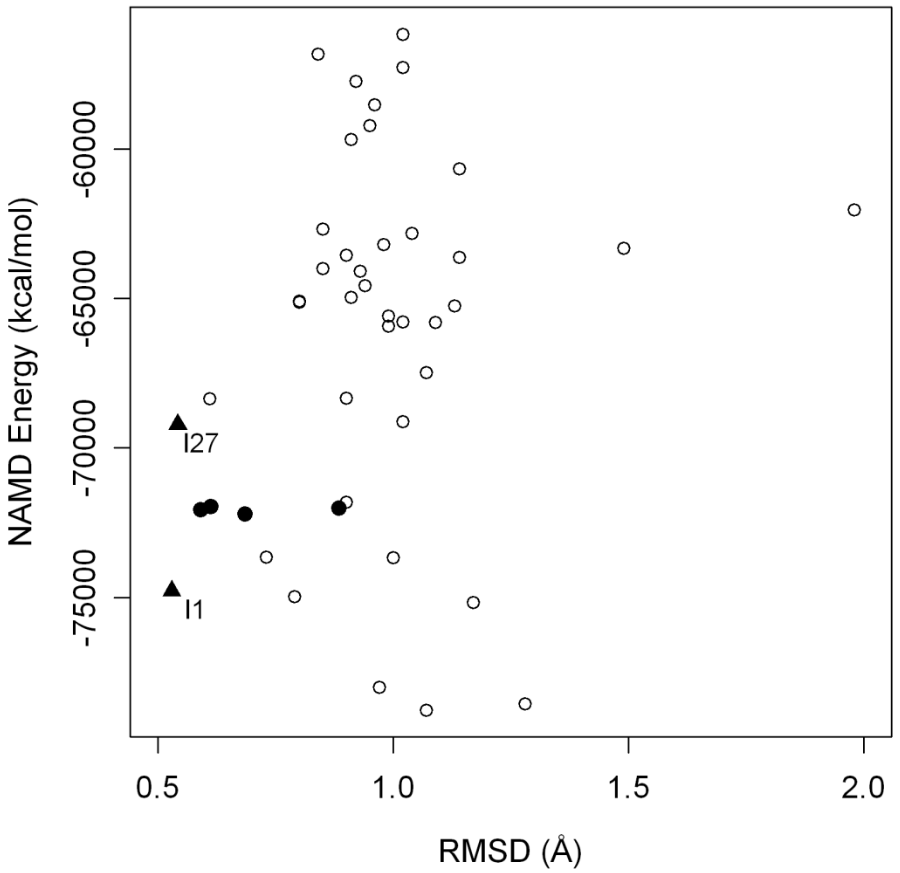


Figure 2.3 Minimized energies of the engineered proteins and wild types.

The RMSD between models and templates were plotted with the total energy after NAMD minimization. For I1 and I27, their crystal structures (PDB: 1G1C and 1WAA) were minimized and shown in filled triangles. The models for the rest 38 Ig-like domains in Titin I-band are shown in empty circles and the models for four engineered proteins are in filled circles.

Adapted from (Lu et al., 2012).

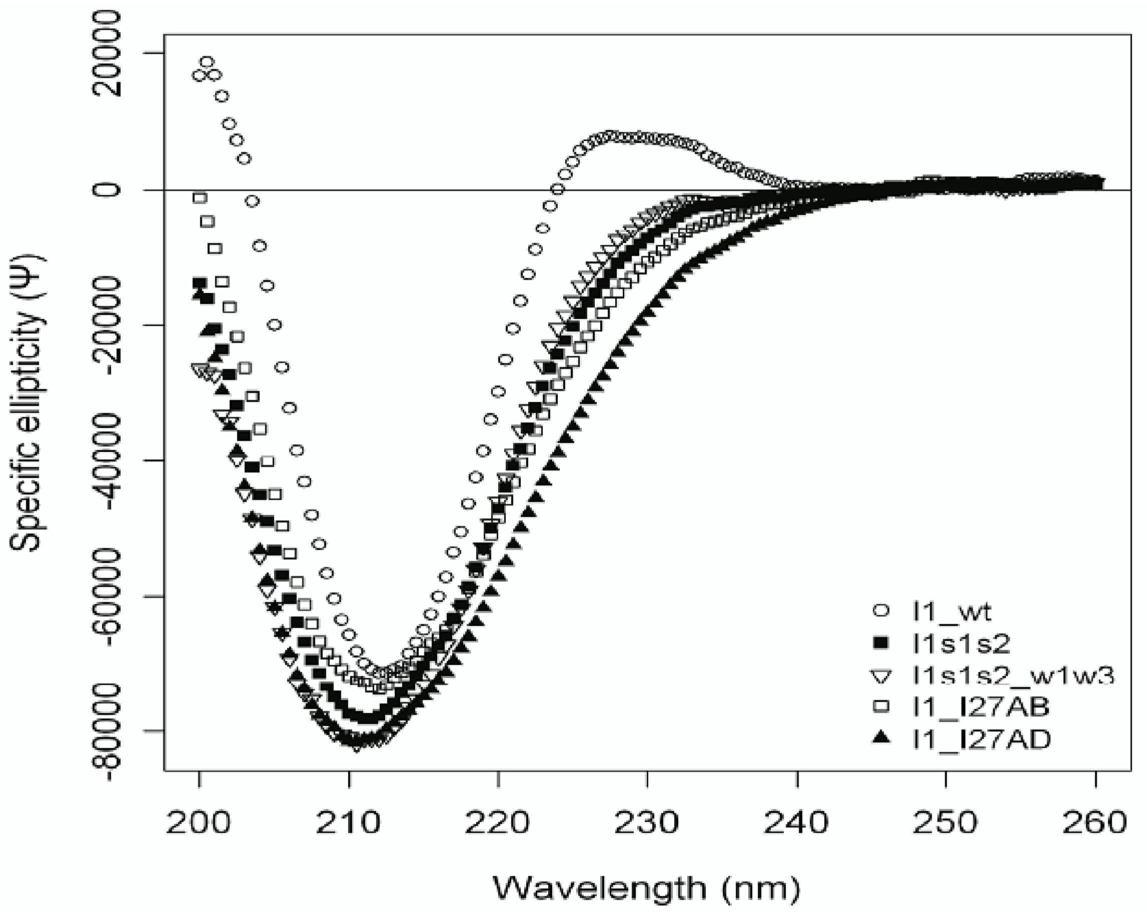


Figure 2.4 Far-UV CD spectra for the engineered proteins.

The spectra of all engineered proteins showed features of predominantly β-strands secondary structures, which is similar to wild type I1.

Adapted from (Lu et al., 2012).

The mechanical strengths of the engineered novel Titin domains

The mechanical strengths of the designed domains were measured by AFM in the context of a poly-protein with five domains (Figure 2.5 A). The target domain, which may be either one of the four designed domains or wile type I1 (as a control), is located in the middle with two wild type I27 domains on each side. The wt I27 domain serves as a calibration domain for the force measurement since its mechanical strength has been well studied (Li et al., 2000). The molecular weight of the poly-protein is ~50KD, which was verified by SDS-PAGE (Figure 2.5 B).

With wild type I1 as target domain, we observed one low peak for the I1 domain and four high peaks for the I27 domains in the force extension profile as expected (Figure 2.5 C). The four peaks around 200pN correspond to the force needed to unfold the four I27 domains and the lower peak around 100pN corresponds to the unfolding of wild type I1 domain. Statistical analysis of the peaks in 44 recordings confirmed this assignment with a mean force of ~100pN (98 ± 22pN) in 44 low peaks and ~200pN (201 ± 27 pN) for 176 high peaks (Figure 2.5 D). The bimodal distribution of these unfolding forces demonstrates the significant difference of mechanical strengths between I1 and I27 domains.

For the poly-protein containing the I1s1s2 target domain, we observed 5 peaks with approximately same unfolding force (Figure 2.6 A). As one of these peaks must correspond to the unfolding of the target domain, the designed domain I1s1s2 has approximately the same strength of 200 pN as the wild type I27. Thus introducing two strong motifs in I1 increased the mechanical strength of the designed domain by about 100pN. This finding is further verified by the unimodal distribution of the unfolding forces in all 216 peaks from repeated AFM experiments (Figure 2.6 B). The average value of 204 pN for the unfolding forces coincides with the expected value for wt I27. The force-extension curves observed for the poly-protein with I1s1s2-w1w3 as target domain similarly show four or five peaks with maxima values close to that of the wild type I27 (Figure 2.6 C) and the histogram of the unfolding forces in 156 experiments is unimodal as for the I1s1s2 domain, albeit slightly shifted to lower unfolding forces with an average value of 187 pN (Figure 2.6 D). This observation indicates that the mechanical strength of the designed domain I1s1s2-w1w3 increased by about 90pN. In the case of the I1_I27AB poly-protein we found that many of the force-extension curves show a low peak of about 100 pN (Figure 2.6E). The histogram of the force barriers found in the recordings of 206 experiments shows a bimodal distribution with many unfolding force peaks near 100pN (Figure 2.6 F). Hence the I1_I27AB mutant does not change the mechanical stability of I1. We were not able to obtain clean records for the poly-protein containing the domain I1_I27AD as target, probably due to the low stability of that domain. The AFM results for the designed domains I1s1s2 and I1s1s2-w1w3 clearly demonstrate the success of the design strategy and encourage us to further validate this strategy with other domain structures and other folds.

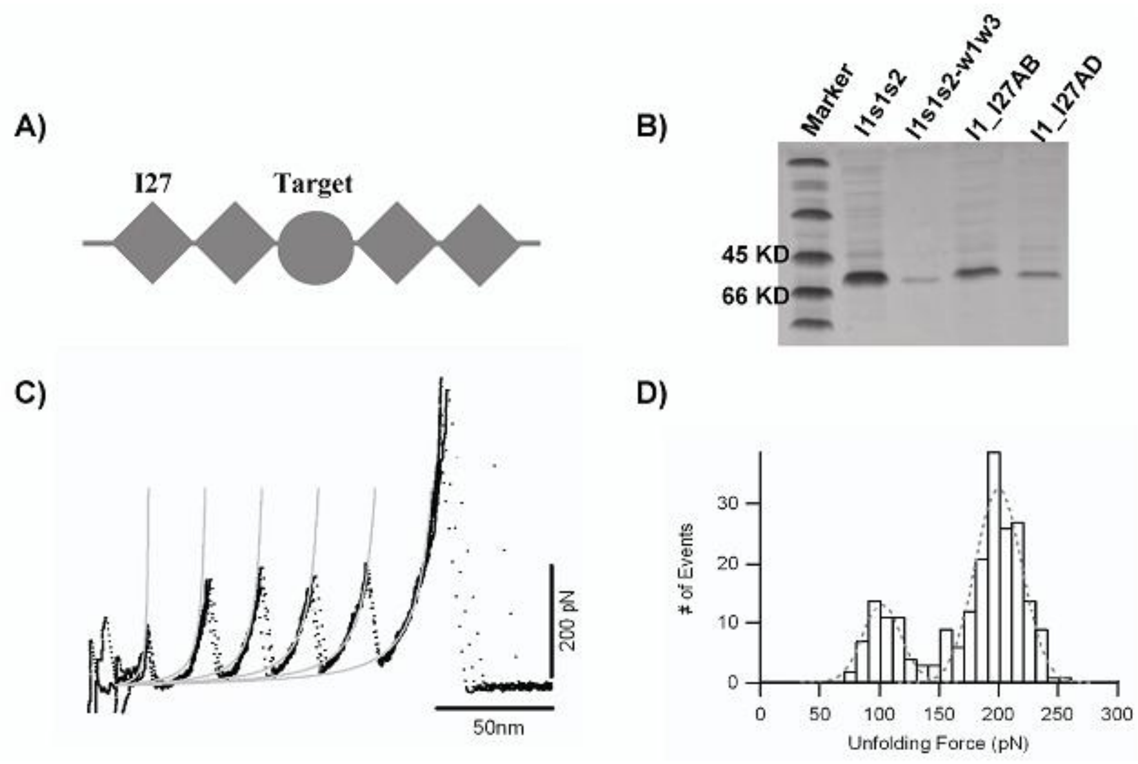


Figure 2.5 Poly-protein and Force-extension profile of wild type I1.

A), the poly protein consists of four I27 domains (diamond) and one target domain (circle). The target domain could be wild type I1 or engineered proteins. B), the SDS-PAGE shows the MW of the poly proteins are ~50KD. C), the poly protein containing wild type I1 domain was used as a control. The force–extension curve shows four peaks for I27 and one low peak for I1. Four curves were overlaid in the figure. Grey curves show the fitting of worm-like chain. D), the histogram of unfolding forces shows two separate peaks. The peak at 98±22 pN, n=44 corresponds to wild type I1, and wild type I27 unfolds at 201±27 pN, n=176. Dashed lines are Gaussian fit to the histogram.

Adapted from (Lu et al., 2012).

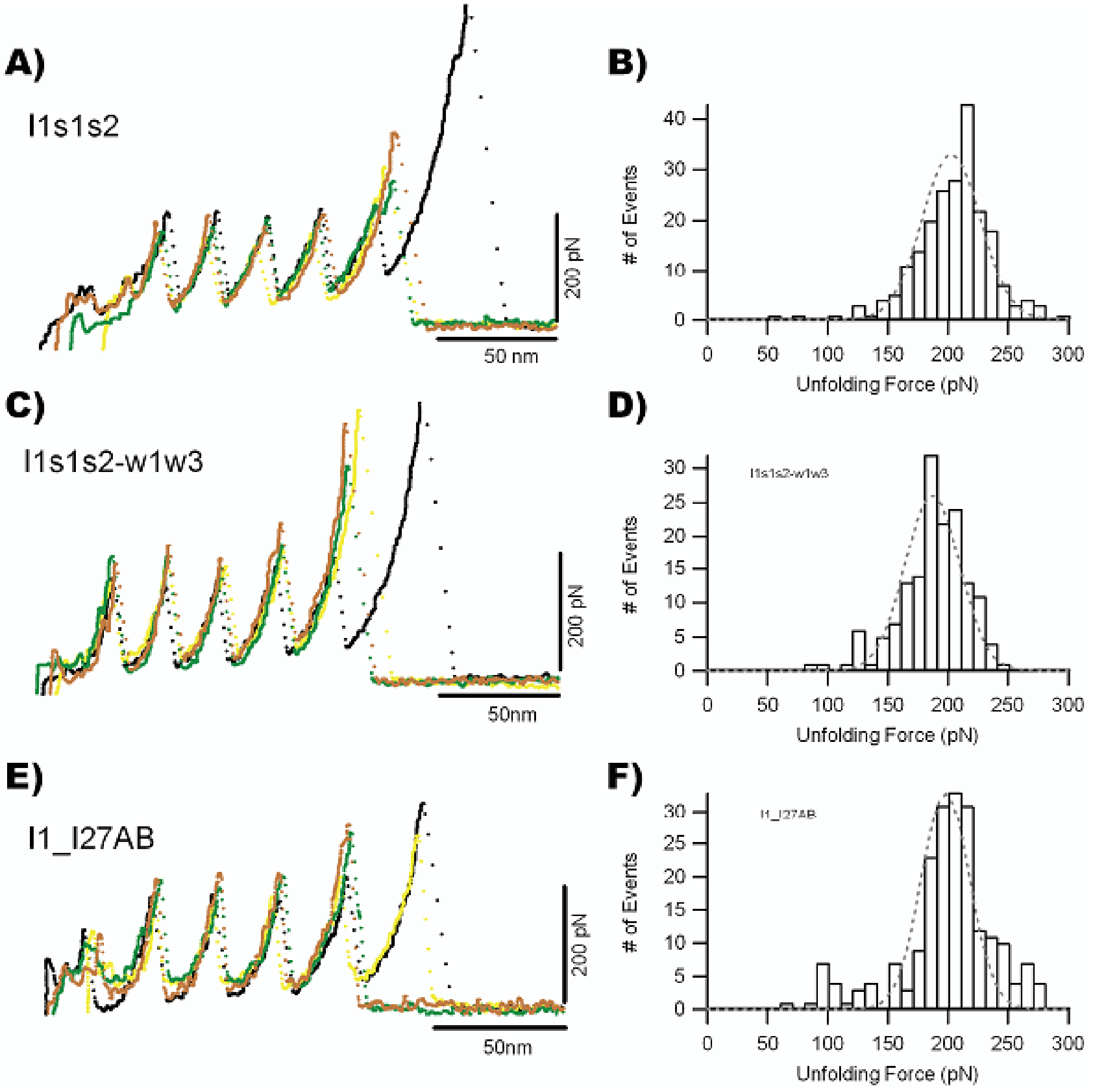


Figure 2.6 Force-extension profiles of engineered proteins.

Four saw-tooth curves were overlaid for each engineered protein and shown in different colors (black, green, red and yellow). Average unfolding forces and standard deviations were calculated according to the Gaussian fits. A), B), there are four or five peaks with almost equal height. One of them must correspond to the unraveling of the engineered protein I1s1s2. The distribution of unfolding forces shows only one peak around 200pN. (204±32 pN, n=216) C), D), Similar to I1s1s2, the extension profile of I1s1s2-w1w3 also shows four or five equal peaks; however, the peak of the histogram is shifted to left. (187±28 pN, n=156) E), F), many of the extension profiles for I1_I27AB show a small peak at the beginning. The histogram also has a shoulder at ~100pN (198±40 pN, n=206). Adapted from (Lu et al., 2012).

Steered Molecular Dynamics simulation of the engineered proteins

SMD simulations have been successful applied in the analysis of the unfolding of proteins under external force. The simulation attempts to mimic the forced unfolding process in AFM experiments by fixing one end of the protein while applying external forces on the other end. However, due to the limited computing power, the pulling velocity in simulation is usually 10^5-10^8 higher than in experiments to ensure the simulation finishes within a reasonable period of time, which requires much higher force to unfold the protein than in experiments. Even with this limitation, the resulting force-extension profiles and unfolding trajectories of SMD simulation still provide insights into the detailed events during the rupture of Ig domains.

The mechanical stability of recombinant I1s1s2 was shown to be enhanced in AFM experiments. Constant velocity SMD simulations were finished at different velocities (0.5, 0.1 and 0.01Å/ps) with the homology model of I1s1s2 to analyze the contribution of the introduced strong motifs to mechanical stability. Unfolding of the X-ray structures of I1 and I27 (PDB: 1G1C and 1WAA) were also simulated as control. Previous results have shown that the force-bearing mechanical “clamp” of A-B, A’G strands breaks between extension of 10-20 Å, thus, the simulations were continued until the extension of C terminus reaches 25Å when the clamp is fully broken. While the force-extension profiles do not show much difference between I1s1s2 and I1, I27 at higher velocities, at pulling speed of 0.01Å/ps, the unfolding forces of I1s1s2 are very similar to I27 and higher than I1 between the extension of 13-20 Å, suggesting that the strong motifs stabilize the force-bearing regions (Figure 2.7 A, orange box). In addition, the interaction energies between each pair of residues involved in the force bearing regions or strong motifs in I1s1s2 were monitored to identify the residue pairs whose interaction energies are enhanced due to the introduction of motifs. For example, the interaction forces between residue 6 in strand A and residues 22-26 in strong motif 2 were recorded during the simulated unfolding of I1s1s2. The force-extension profiles indicate that the interactions between residue 6 and 22, 24 remain strong during the rupture of force-bearing regions (Figure 2.8 A, red, blue). Many other residue pairs that are strongly interacting with each other during unfolding were identified and mapped on the structural model of I1s1s2 (Table 2.1, Figure 2.8 B, yellow). The introduced motifs enhanced the interactions between these residue pairs and thus increased the mechanical stability of I1s1s2.

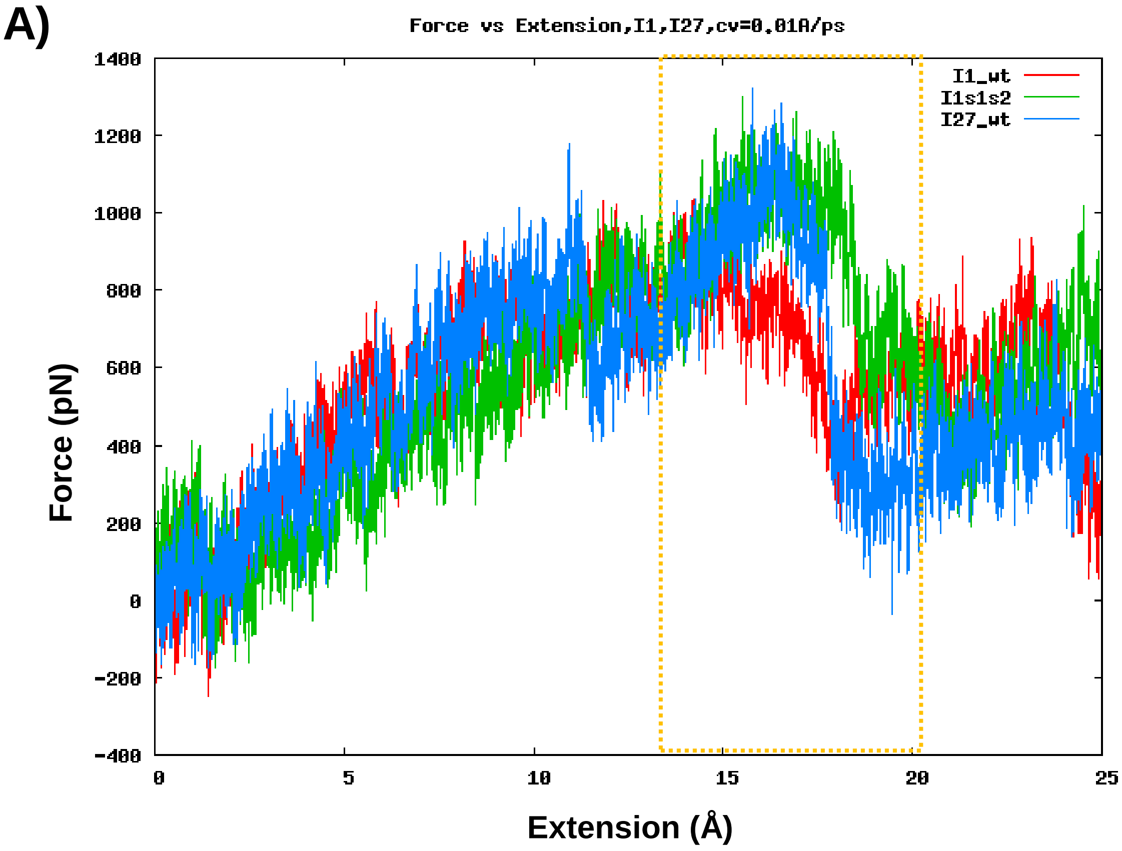


Figure 2.7 Force-extension profiles of SMD simulation

Red: I1 wild type; Green: I1s1s2; Blue: I27 wild type. Orange box: important events in the rupture of force-bearing regions.

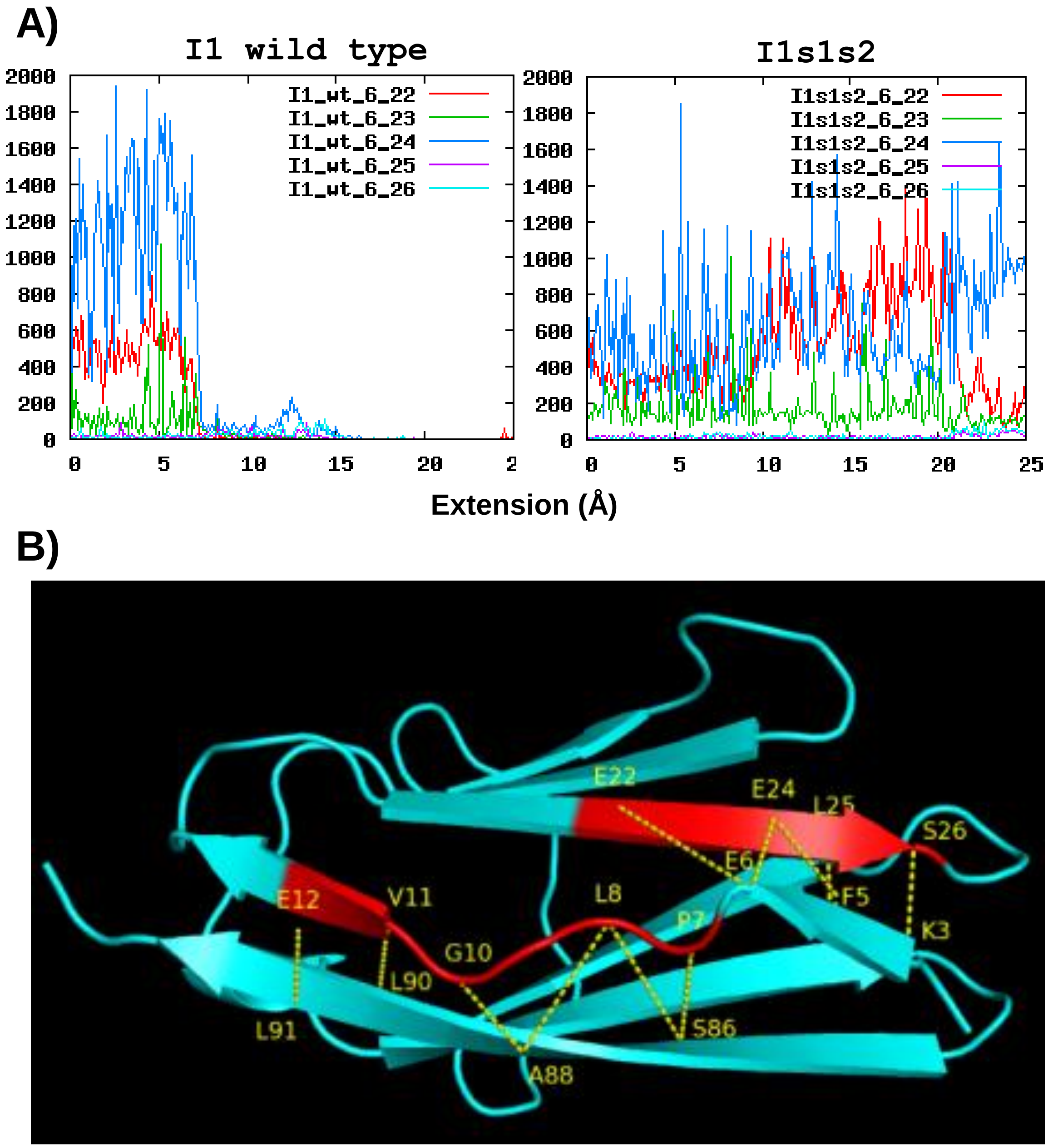


Figure 2.8 Pairwise interactions in SMD simulation

A). pairwise force-extension profiles of residue 6 and 22-26.

B). residue pairs that have strong interactions are mapped on I1s1s2

Table 2.1 Residues involved in pairwise interactions

Location of residues	Pairwise interactions		Location of residues
Strand A	K3	S26	Strong motif 2, Strand B
Strand A	I4	L25	Strong motif 2, Strand B
Strand A	F5	E24	Strong motif 2, Strand B
Strand A	E6	E22	Strong motif 2, Strand B
Strand A	E6	E24	Strong motif 2, Strand B
Strong motif 1	P7	S86	Strand G
Strong motif 1	L8	S86	Strand G
Strong motif 1	L8	A88	Strand G
Strong motif 1	G10	A88	Strand G
Strong motif 1, Strand A’	V11	L90	Strand G
Strong motif 1, Strand A’	E12	L91	Strand G

2.4 Discussion

The mechanical stability is an important property for many structural proteins involved in cell adhesion, translocation and contraction, such as the skeletal and cardiac striated muscle, connective tissue, and various kinds of epithelia (Granzier and Labeit, 2004; Kruger and Linke, 2009; Linke, 2008). The mechanical strength of many proteins with different folds has been characterized by single molecular AFM experiments (Sikora et al., 2011). Promising progress has been achieved to understand the mechanisms of mechanical stability by additional analysis with computational simulations. Several structural motifs, such as the “cysteine slipknot” (Sikora et al., 2009; Sulkowska et al., 2010) and “mechanical clamp” (Gao et al., 2002a; Gao et al., 2002b; Hsin et al., 2011) have been revealed. However, the rational design of proteins with specified mechanical properties remains challenging (Crampton and Brockwell, 2010; Galera-Prat et al., 2010), and current strategies resemble more a trial and error approach rather than a rational procedure.

SMD calculations of the unfolding process of the β-sandwich domains (Ig-like domains, fibronectin domains) have provided microscopic details and indicated the importance of hydrogen bonds between A-B, A'-G strands in maintaining the mechanical stability (Gao et al., 2002a; Gao et al., 2002b; Hsin et al., 2011; Lu and Schulten, 2000). However, it is still an open question to what extent other interactions, such as side chain packing and the hydrophobic interactions of the core are involved in the mechanical stability. A disadvantage of the all atom simulations is that most previous calculations were conducted at very high pulling velocities that are about six orders of magnitude higher than experiments due to insufficient computing power, resulting in unrealistic unfolding forces that are much higher than AFM measurements. Only few recent studies were able to simulate the unfolding forces at the magnitude of AFM measurements either by simulating at lower velocities (491 pN for Titin I91 at 0.028 Å/ns) (Lee et al., 2009) or using modified SMD strategies, such as PUFF (213pN for Titin I27 at 2.6 Å /ps) (Ho and Agard, 2010). Other simulation methods like the coarse-grained Go-like models (Faisca et al., 2010; Go and Abe, 1981; Sulkowska and Cieplak, 2008) are not capable of providing the same level of details as the all atom simulations. There are also studies aiming to conduct AFM measurement at high speed to enable the direct comparison of unfolding forces between experiments and simulations. Rico et al. developed the high-speed force spectroscopy (HS-FS) and their measurements on Titin I91 at pulling velocity of close to 0.04 Å/ns (comparing to conventional experimental velocities at ~5×10^{-6} Å/ns) showed unfolding forces greater than 500pN, which overlapped with simulation results (Rico et al., 2013). It should be noted that the details revealed by simulations may not completely agree with the forced unfolding of proteins and it should be combined with other analysis for more comprehensive understanding.

In my study, I used Titin as an experimental model system to develop the strategy of engineering proteins with enhanced mechanical strength. Based on previous work on linear sequence motifs of weak and strong Titin domains (Garcia et al., 2009), I demonstrated the usefulness of the PCP motifs for the design of mechanically stable proteins with clear experimental evidences by the AFM recordings. This dramatic increase of the mechanical strength (from ~100pN to ~200pN) is remarkable as studies with similar Titin domains were only partially successful (Crampton and Brockwell, 2010). Rather than claiming that our approach solves the rational design problem, the encouraging experimental results of our design demonstrate the potential of the PCP motifs. We suggest that the method might be most useful in practice in combination with SMD simulations (Hsin et al., 2011; Lee et al., 2009), analysis of side-chain packing in the hydrophobic core (Sadler et al., 2009) and phi-value analysis of the transition state (Best et al., 2003; Borgia et al., 2008). As protein engineering based on those approaches involves a careful selection of the residues for evaluation, our method can be used as an additional filter in the screening of the residues.

2.5 Conclusions and future studies

In this section, I described the results of a first experimental study to evaluate the usefulness of PCP motifs in the design of mechanically stronger domains. The mechanical strengths of two of the engineered proteins have significantly increased, demonstrating the potential of our method. Although several open questions regarding the optimal strategy for combining motifs and the usefulness of weak motifs remain, we suggest that the method can used in combination with other methods as a filter to select relevant residues. The PCP motifs could explain the experimentally observed hierarchy of mechanical strength of all Titin Ig domains, and can be similarly applied to other types of protein folds.

In future studies, we plan to conduct SMD simulations at lower pulling velocities to better mimic the unfolding process of Ig domains and provide more reliable details of unfolding. In addition to the liner force-specific PCP motifs, we aim to identify potential structural motifs that are critical for mechanical stabilities by searching for PCP-conserved residues that are within certain spatial distances.

CHAPTER 3: PREDICTION OF ALLERGENICITY

The prevalence of allergic diseases has been rising dramatically in the US and worldwide in the last few decades (WAO, 2011; Weinberg, 2011). The diseases cause mild symptoms like itchiness, rhinitis, asthma and rash or in severe cases can lead to anaphylactic shocks. There are various sources of allergens including pollen, dust, insect, mold, drug and most importantly, food. Thus, guidelines to distinguish allergenic proteins from others are in urgent need for regulatory agencies, biotech companies and physicians. However, the current criteria suggested by FAO and WHO generate many false positives (FAO/WHO, 2001; FAO/WHO, 2003; Ivanciuc et al., 2009c; Schein et al., 2007). In this project, I aimed to develop bioinformatics tools to evaluate the allergenicity of proteins. Previously the Braun laboratory has shown that allergenic sequences cluster in a small number of protein families (Ivanciuc et al., 2009a). I generated allergen-specific motifs for the major protein families containing allergens. The motifs were used to score and evaluate allergenicity of query proteins. A set of non-allergenic sequences from Uniprot database were used to validate the scoring method and the distribution of scores shows a clear separation from allergenic sequences. In addition, our method also has a lower false positive rate than the FAO/WHO rules. The results suggested that it is necessary to improve the FAO/WHO rules and allergen-specific motifs are very helpful for a better prediction of allergenicity.

3.1 Introduction

Allergenic proteins

Allergenic proteins are usually harmless to the human body, but for certain individuals, they induce vigorous responses of immune system (type-I hypersensitivity reaction) that lead to mild or fatal symptoms. Most allergic reactions are mediated by specific immunoglobulin E (IgE) antibodies and the symptoms occur within minutes to a few hours after exposure to allergens (Kindt et al., 2007). The symptoms range from itchiness, asthma, rash to severe anaphylaxis that require immediate treatment. Multiple studies have reported that the prevalence of allergic diseases is on a rising trend world widely for the last few decades, especially in industrialized nations (Bloomfield et al., 2006; Platts-Mills et al., 2005; WAO, 2011). In the US, statistics from American Academy of Allergy Asthma & Immunology indicate that allergic diseases with mild symptoms like asthma, rhinitis, and sinusitis affect millions of people, and severe anaphylaxis by drug and food allergies cause hundreds of deaths every year (AAAAI; Pleis et al., 2010).

Allergens can be found in various sources in the environment and our daily life, including pollen, dust, insects, food and drugs. Among them, food is one of the major sources of allergens that cause severe allergic reactions. About 6% of US children under age 3 and 3.5–4% of the overall US population are affected by food allergens according to the 2011-2012 white book published by World Allergy Organization(WAO) (Gupta et al., 2011; WAO, 2011). Some food sources, such as peanuts, can be the cause of anaphylactic shocks that lead, in extreme cases, to fatalities of sensitive individuals. Each year there are more than 100,000 hospital visits and 100 to 200 fatal reactions due to food allergy (Sicherer, 2011; Sicherer and Sampson, 2010). The US Food and Drug Administration

(FDA) has established a list of eight most common allergenic foods (Milk, Eggs, Fish, Crustacean shellfish, Tree nuts, Peanuts, Wheat and Soybeans) which account for 90% of food allergic reactions and the labeling of using these ingredients are enforced (FDA, 2004). It is also well established that allergens may be cross-reactive with other allergens (Bonds et al., 2008; Schein et al., 2007). Some individuals who are sensitive to certain allergens may develop allergic reactions to other foods that contain similar proteins to the known allergens. For example, those allergic to the major birch pollen allergen Bet v 1 may develop a secondary allergy to the similar protein in peanut, Ara h 8 (Mittag et al., 2004). Cross-reactivity is usually due to similar sequences, structures or homologous proteins, however, it is still unclear what the exact causes are and it is a challenging task to predict or prevent cross-reactivity.

Genetic engineering is one of the most powerful technologies that have been increasingly applied in modern agriculture for the last two decades. The technology has significantly contributed to food production and related industries under the pressure of rapidly growing population, loss of arable land and agricultural pests and disease (Goodman and Tetteh, 2011). Cultivation of commercial genetically modified crops began with approximately 4.3 million acres globally in middle 90s, and by 2010, about 15.4 million farmers grew approved modified crops on 366 million acres in 29 countries, representing 81% of soybean, 64% of cotton, 29% of corn, and 23% of canola cultivation (ISAAA, 2010). Genetic modification of crops introduces foreign genes from other organisms into the genome of recipient plants to enable features like pest-resistance or herbicide-tolerance. However, despite the benefits of genetic modification, insertions of foreign genes and their metabolites also raise the concern of increasing potential risks of allergenicity in those foods and related products. Thus, assessment of the risk of food safety is strongly suggested to help regulators and companies to identify potential hazard and minimize the chance of allergenicity and toxicity (Delaney et al., 2008; Goodman et al., 2008; Ivanciuc et al., 2009c).

Assessment of allergenicity

In silico bioinformatics tools like FASTA (Pearson and Lipman, 1988) and BLAST (Altschul et al., 1990) are applied to the assessment to provide fast and inexpensive initial evaluations. The Food and Agriculture Organization (FAO) and the World Health Organization (WHO) recommended two criteria for the evaluation of allergenicity by comparing a query protein with a known allergen: 1) more than 35 % sequence identity using a window of 80 amino acids and a suitable gap penalty; 2) identity of 6 contiguous amino acids (FAO/WHO, 2001; FAO/WHO, 2003). Proteins that are determined positive using these two rules are considered to be cross-reactive to known allergens and need further experimental examination by testing the binding of the query protein to IgE from patients who are sensitive to that allergen. While the criteria performs quite well on identifying known allergens with sensitivity of >90%, the main pitfall is that both rules tend to over-predict cross-reactivity between query proteins and allergens and produce large amount of false positives (Bjorklund et al., 2005; Gendel, 2002; Goodman, 2006; Hileman et al., 2002; Kleter and Peijnenburg, 2002; Li et al., 2004; Silvanovich et al., 2006; Stadler and Stadler, 2003) (Figure 3.2). This causes great burdens for the experimental validation which are rather expensive, time consuming and difficult to interpret (Goodman and Leach, 2004). There have been many suggestions on modifying these guiding criteria by increasing the cutoff value of 35% sequence identity and 6 exact-match residues (Goodman, 2008; Ladics et al., 2007).

While the criteria of FAO/WHO mainly emphasize similarity of sequences and exact match of short peptides, many other bioinformatics tools have been developed based on either classification algorithm or sequence profiles/motifs (Cui et al., 2007; Ivanciuc et al., 2009b; Martinez Barrio et al., 2007; Muh et al., 2009; Saha and Raghava, 2006; Soeria-Atmadja et al., 2006; Stadler and Stadler, 2003; Wang et al., 2013). Most of the classification algorithms like Support Vector Machine (SVM) still use composition of amino acids or allergen-relevant peptides as input vectors. Methods using sequence profiles rely on rather unspecific motifs and limited number of known IgE epitopes. In our research, we approached the evaluation of allergenicity by scoring query sequences against allergen-specific motifs. With the collection of allergenic sequences in our database SDAP (Ivanciuc et al., 2003), we were able to generate allergen-specific motifs for the 17 Pfam-A families that contain major allergens. Our scoring method demonstrated its ability to separate allergenic and non-allergenic sequences even when they are homologues. The method can also be used to imply potential cross-reactivity between different allergens.

3.2 Methods

SDAP

The Structural Database of Allergenic Proteins (SDAP) is a database maintained by the Braun laboratory that includes the sequences and 3D structures for allergenic proteins

(Ivanciuc et al., 2003). It is available to the public at https://fermi.utmb.edu/SDAP/ and frequently used by researchers interested in allergenic proteins. In SDAP, experimental IgE-binding epitopes and experimental 3D structures of allergens are collected by reviewing recent literatures. For the allergenic proteins without 3D structures, great effort has spent in generating high quality homology models for them and the models were shown to be reliable (Oezguen et al., 2008b; Power et al., 2013). Various computational tools are also integrated in the database to assist the sequential and structural analysis related to allergens (Ivanciuc et al., 2009b). For example, a web server is developed to allow the evaluation of allergenicity of protein sequences using the FAO/WHO guidelines (https://fermi.utmb.edu/SDAP/sdap_who.html). SDAP is a powerful tool in the investigation of the cross-reactivity between known allergens and predicting the potential IgE-binding epitopes. The allergens are cross-referenced to other common protein sequence and structure databases such as SwissProt, PIR, NCBI, PDB and PFAM. Currently there are 1273 allergenic protein sequences of 1526 Allergens and isoallergens, 92 allergens with PDB structures, 458 allergens with 3D models and 29 allergens with IgE epitope sets in SDAP.

Datasets

A collection of allergenic sequences is needed for generating allergenic motifs. I searched through all 1273 sequences in SDAP and identified 17 Pfam-A families which contain more than 10 allergenic sequences. The sequences that belong to these families were collected and grouped according to their Pfam families. Since there are many iso-allergens which share highly similar sequences, and the redundancy would bias the sequence motifs, I used CD-HIT (Li and Godzik, 2006) to cluster the sequences for each family and selected representative sequences at the cutoff of 95% sequence identity. Finally 568 sequences were used to generate motifs. Then multiple sequence alignments were prepared with ClustalW (Larkin et al., 2007) and sequence motifs were generated using PCPMer for each Pfam family.

Another set of non-allergenic sequences is necessary for validation of the allergenic motifs. We obtained a large set of protein sequences from UniProtKB (Magrane and Consortium, 2011) and reduced redundancy to 40% sequence identity using CD-HIT. Then we removed any potential allergenic protein if the sequence meets any of the following criteria: 1). the description of the sequence contains any keyword related to allergen (allergen, allergy, lipid transfer protein, profilin, lipocalin, pectate lyase, tropomyosin, melittin, thaumatin, seed storage protein); 2). the sequence belongs to any of the protein families contain allergenic sequences, like PF00235 or PF00407. Sequences that are shorter than 50 amino acids were also removed. There are 84939 sequences in the final set of non-allergenic sequences.

Scoring method

The algorithm of PCPMer to generate sequence motifs is described in Chapter 1 (Mathura et al., 2003). In general, PCPMer identifies the PCP profiles of conserved motifs for a given set of sequences. Another powerful feature of PCPMer is to evaluate query sequences using given motif profiles and identify sequences containing similar profiles. Sequences with similar motif profiles are likely to share similar function or properties. In this project, we used the motif profiles generated for the allergenic sequences from 17 Pfam-A families as criteria and developed a scoring method to evaluate the potential allergenicity of query sequences.

To determine how well a motif matches a sequence, the motif is aligned to every position *k* of the sequence and the score values *S(k)* are calculated based on the Lorentzian scores scheme. For each position *i* within the motif, the score

$$S_{k,i}^{p} = \left[1 + \left(\frac{V_{k+i}^{p} - \langle V_{i}^{p} \rangle}{W\sigma_{i}^{p} + \Phi}\right)^{2}\right]^{-1}$$

Here, *k* denotes the position in the query sequence, to which the motif has been aligned; *i* is the position index within the motif of the length n, and it loops from *0* to *n-1* ; V_{k+i}^{p}, *p* are the five quantitative descriptors (E1-E5) of the amino acid at the position $k + i$ in the sequence; $\langle V_{i}^{p} \rangle$ is the average PCP value at the position *i* of the motif and σ_{i}^{p} is the corresponding standard deviation. *W* is the weight for standard deviation (by default set to 1.5) and the small positive shift Φ (set to 0.001) was added to prevent overflow during calculation when standard deviation is zero.

The score value for the motif of the length *n* aligned at the position *k* in the sequence is then calculated as the average of the scores at each position *i* of the motif.

$$S(k) = \frac{1}{5n} \sum_{\substack{p=E1..E5 \\ i=0..n-1}} S_{k,i}^{p}$$

The resulting score *S* of a motif against query sequence is the maximum of the score values calculated for each position k, *S(k)* . The scores S are values from the interval 0 to 1, with 1 indicating a perfect match.

Usually more than one motif will be generated from a multiple sequence alignment. A total score is calculated to evaluate if a query sequence matches a set of *m* motifs with the scores S_1*...Sm*

$$S_{tot} = \sum_{i=1}^{m} \frac{S_i |\langle S \rangle_{aln} - \langle S \rangle_{db}|}{(\sigma_{aln} + \sigma_{db})^2 + \epsilon}$$

Where i loop through each motif; $\langle S \rangle_{aln}$ and σ_{aln} are the mean and standard deviation of scores in the multiple sequence alignment that was used for generating the motifs; $\langle S \rangle_{db}$ and σ_{db} are the mean and standard deviation of scores in the query set of sequences. If there is only one query sequence, $\sigma_{db} = 0$. The small positive shift (by default set to 10^{-10}) is added to prevent division by zero.

To evaluate the potential allergenicity of proteins, the motifs for the allergenic sequences from the 17 Pfam-A families were used as criteria for the scoring. A query sequence was scored against all 17 sets of motifs using the method described above, resulting in 17 scores $S_{tot,i}$, where $i = 1 \ldots 17$. Since the scores of the query sequence are based on motifs from different Pfam-A families, $\langle S \rangle_{aln}$ and σ_{aln} are different between different Pfam-A families, which means that $S_{tot,i}$ of the query sequence are not on the same scale. In order to compare the scores of the query sequence, I used the set of non-allergenic sequences from UniProt as background and calculated the total scores $S_{tot,i}^{b}$ for each protein family. Then the scores of query sequence were converted to standard scores z_i according to the background score distribution of non-allergenic sequences $S_{tot,i}^{b}$ for each family.

$$z_i = \frac{S_{tot,i} - \langle S_{tot,i}^{b} \rangle}{\sigma_{tot,i}^{b}}$$

Finally, the highest z score of the query sequence was taken as the result to indicate the extent of allergenicity.

$$z = \max(z_i)$$

Binding efficiency test

To validate the significance of the motifs identified by PCPMer, we tested the binding of peptides matching these motifs to IgE from patient sera in collaborations with Drs. C. H. Schein and S. J. Maleki (USDA, New Orleans) (Figure 3.1). Peptide microarray assays are powerful tools to study the binding efficiency for many different peptides. The motifs were synthesized and spotted onto glass microarray slides by JPT Peptide Technologies (JPT Peptide Technologies GmbH (Berlin, Germany). For assay, slides were placed in the slide holders within the hybridization chamber of the HS 400 Pro (Tecan, Austria, GmbH, Salzburg, Austria). The Blocking solution, serum, and other solutions were injected into each slide chamber through the injection port for each slide chamber. Slides were blocked for one hour at 4 Celsius degrees washed three times and incubated with individual patients' sera. The serum from individual patients allergic to peanuts and/or tree nuts were centrifuged and the supernatant was injected into the injection port of the HS 400 hybridization chamber on to each array slide and incubated at4℃ overnight (14-16 hours). The slides were then washed. Next, mouse-anti-human IgE (Southern Biotech, Birmingham, AL) was added and the slides were incubated for 1 hour at room temperature in the dark, then washed as before and incubated with anti-mouse IgG Cy3 (green dye read at 532 nm) (Invitrogen, Grand Island, NY) for another 1 hour at room temperature in the dark. The slides are then washed and dried by centrifugation prior to scanning with a GenePix-4000B, Software: GenePix-Pro6 Scanner. The binding of patient’s IgE for the peptides was measured by the fluorescence signal of Cy3. The resulted fluorescence intensities were organized in a spreadsheet for later statistical analysis.

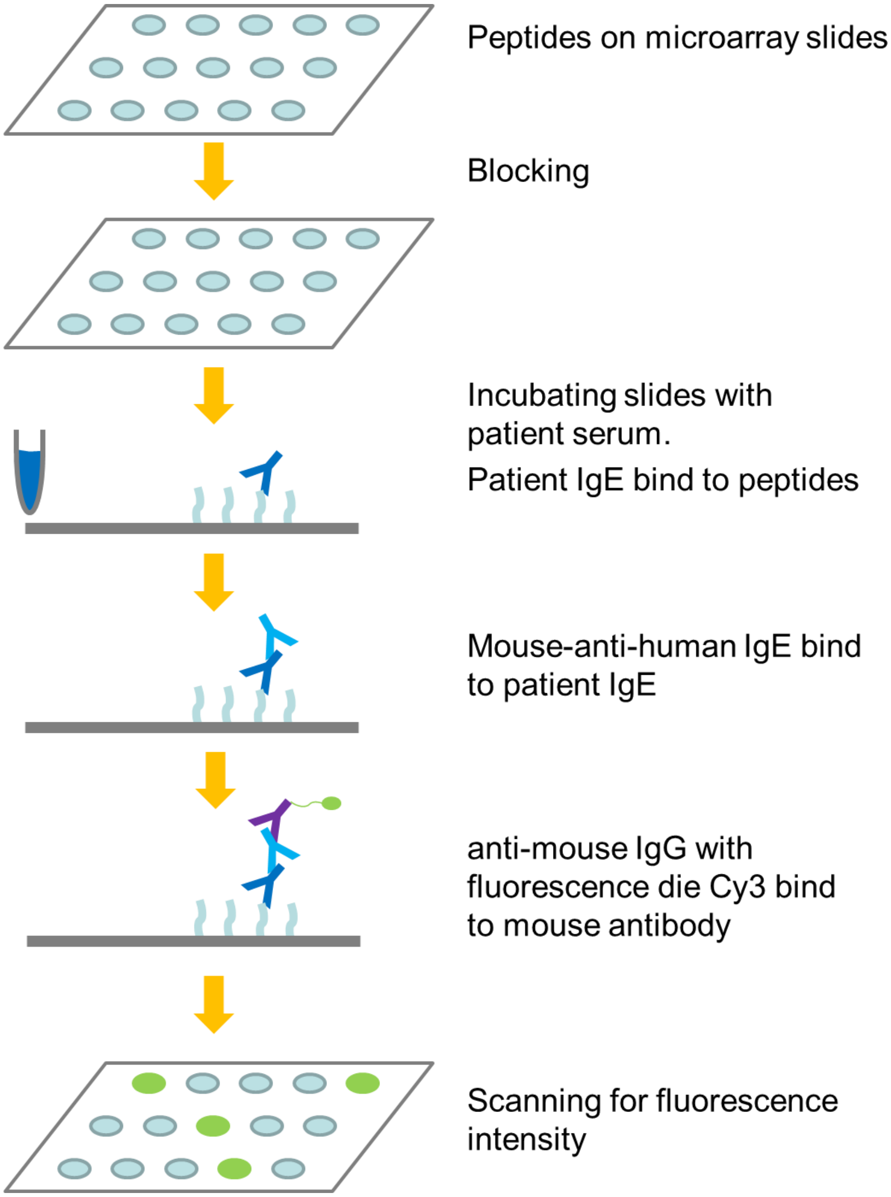

Peptides on microarray slides
Blocking
Incubating slides with patient serum.
Patient IgE bind to peptides
Mouse-anti-human IgE bind to patient IgE
anti-mouse IgG with fluorescence die Cy3 bind to mouse antibody
Scanning for fluorescence intensity

Figure 3.1 Peptide microarray assays

3.3 Results

Evaluation of FAO/WHO criteria

To enable in silico evaluation of allergenicity for query proteins, the FAO and WHO recommended two criteria based on sequence comparison (FAO/WHO, 2001). However, there have been several critics on these criteria suggesting that they over-predict allergenicity and generate too many false positive results, especially the criteria of exact match of 6 amino acids (Goodman, 2008; Herman et al., 2009; Hileman et al., 2002). To verify the rate of false positive, we randomly selected 1000 sequences from a total number of 84939 non-allergenic sequences from Uniprot, and for every window of 80 amino acid of each sequences, we used FASTA (Pearson and Lipman, 1988) to identify matches in allergenic sequences from SDAP. The sequence identity of each window and its match were calculated and the sequence was considered as positive if any window has sequence identity above cutoff value. At different cutoff value of sequence identity at 30%, 35%, 40% and 45%, the false positive rates are 57%, 13.9%, 3.9% and 1.4% (Figure 3.2, blue). On the other hand, the criterion of identical peptides has much higher false positive rates of 85.5%, 23.9%, 4.1% and 1.9% at different peptide length of 6, 7, 8 and 9 (Figure 3.2, red). The high false positive rates indicate that the criteria suggested by FAO and WHO over ten years ago should not be considered as the only guidelines for assessment of allergenicity. It is necessary to either update the cutoff values of these criteria or develop more comprehensive methods.

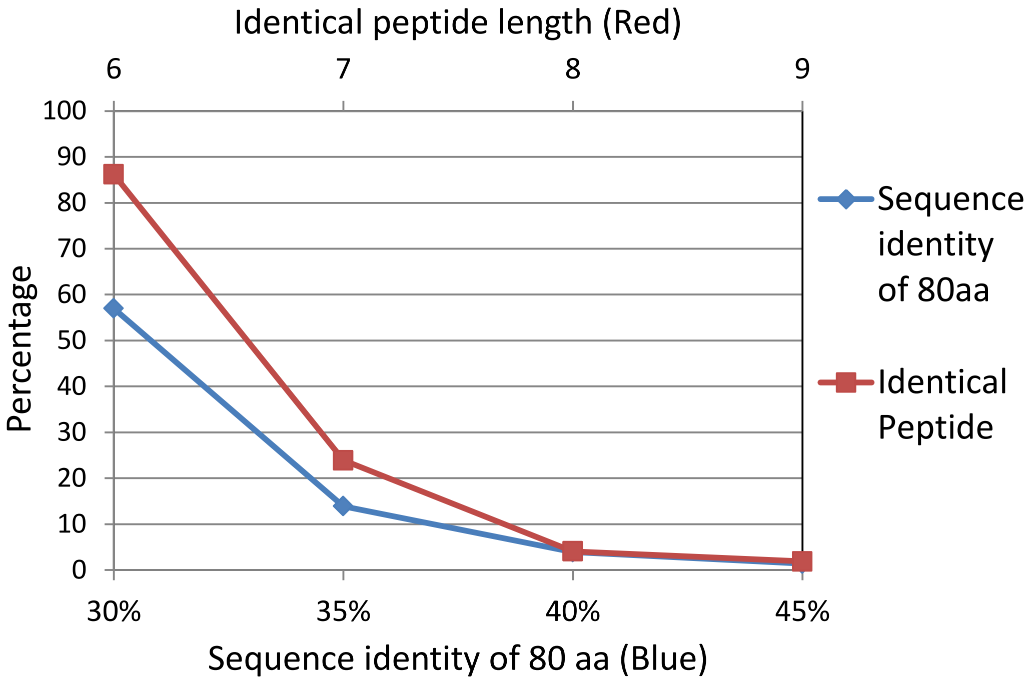


Figure 3.2 Evaluation of the FAO/WHO criteria of allergenicity.

The blue line shows the false positive rate by different value of sequence identity in window of 80 amino acids, and the red line shows the false positive rate by different length of identical peptide (top horizontal axis).

Allergenic motifs for 17 Pfam-A families

We have annotated most of the allergenic sequences in SDAP with their protein families defined by Pfam database (Finn et al., 2014). Even though there are more than 1300 allergenic sequences in SDAP, they are associated with only 130 Pfam-A families and a large portion of these sequences are from a small number of protein families. The aggregation of allergenic sequences suggests that many of them share common features and some of the features may be critical for their allergenicity. In order to characterize the properties of allergens, we identified the proteins families that are abundant with allergenic sequences. As a result, 17 Pfam-A families with more than 10 allergenic sequences reported in SDAP were chosen. The allergenic sequences in these families were collected and grouped by their families. To avoid bias by iso-allergens, we also removed the sequences whose sequence identity to the others is above 95%. Multiple sequence alignments were prepared with ClustalW; and then a set of Physical-Chemical Property (PCP) motifs were generated for allergenic sequences in each Pfam-A family with PCPMer (Mathura et al., 2003). Meanwhile, since some of these motifs may be related to common structural or functional features of all proteins in that protein family whose members include non-allergenic sequences as well, it's necessary to identify this type of motifs. Therefore, we obtained the sequences alignments of these protein families from Pfam database and generated general sequence motifs for each family. The general motifs were removed from the list of motifs generated from allergenic sequences whenever they match each other. The rest of the motifs are considered to be allergen-specific motifs which are unique to the allergens in each protein family and could serve as fingerprints to detect potential allergenic sequences. These 17 Pfam-A families cover all the major types of allergens, especially allergens from various food sources specified by FDA (FDA, 2004), suggesting that our motifs are able to represent the characteristics of the major food allergens. (Table 3.1).

The uniqueness of the allergen-specific motifs to allergens enables us to develop methods to distinguish allergenic sequences from non-allergenic ones. Our scoring method compares query sequences with the motifs and provides a score value as the estimation of potential allergenicity. To validate this method, we generated scores for a set of non-allergenic sequences from Uniprot and allergenic sequences from SDAP (Figure 3.3). The distribution of scores shows a clear separation between non-allergenic sequences and allergenic sequences from the 17 Pfam-A families, indicating the ability of the method to distinguish allergenic sequences from the background. However, the allergenic sequences that are not related to the top 17 Pfam-A families score poorly, mainly due to the lack of characteristic motifs for scoring. This problem could be fixed with the accumulation of new allergenic sequences in those protein families in the future. To estimate the false positive rate, we looked at the percentage of positives at different score values (Figure 3.4). As the cutoff score increases, the positive rates for both allergenic sequences and non-allergenic sequences decrease. The maximum difference of positive rate occurs at score value of 2.7, where above this score, only 4.52% of the non-allergenic sequences are reported as false positive and 92.96% of the allergenic sequences in the top 17 protein families are detected. The false positive rate of our method is lower than the FAO/WHO guidelines at default cutoff of 35% sequence identity (13.9%) or 6 identical amino acids (86.2%), and is similar to the false positive rate of 40% sequence identity (3.9%) and 8 identical amino acids (4.1%) (Figure 3.2).

Table 3.1 Summary of allergens in top 17 Pfam-A families in SDAP

Pfam	No. of allergen	No. of Sequences	Domain	Related Food Sources
PF00234	35	109	Plant lipid transfer proteins	peanut, soybean, tree nut, wheat
PF00235	27	38	Profilin	peanut, soybean, wheat
PF01357	24	47	Pollen allergen	soybean, lobster, wheat
PF00036	23	64	calcium-binding proteins	fish
PF00407	20	114	PR protein, Bet v I family	peanut, soybean
PF00188	19	19	CAP proteins	
PF00190	16	45	Cupin	peanut, soybean, tree nut
PF00261	16	15	Tropomyosin	Crustacean shellfish
PF03330	15	35	Rare lipoprotein A (RlpA)-like double-psi beta-barrel	soybean, lobster, wheat
PF00061	13	16	Lipocalin	domestic cattle
PF01190	10	15	Pollen proteins, Ole e I family	tomato
PF00544	9	17	Pectate lyase	
PF01620	8	21	Ribonuclease (pollen allergen)	barley
PF02221	8	15	MD-2-related lipid-recognition (ML) domain	
PF00187	5	17	Chitin recognition protein	wheat
PF00273	5	20	Serum albumin family	chicken, domestic cattle
PF06757	2	30	Insect allergen related repeat, nitrile-specifier detoxification	

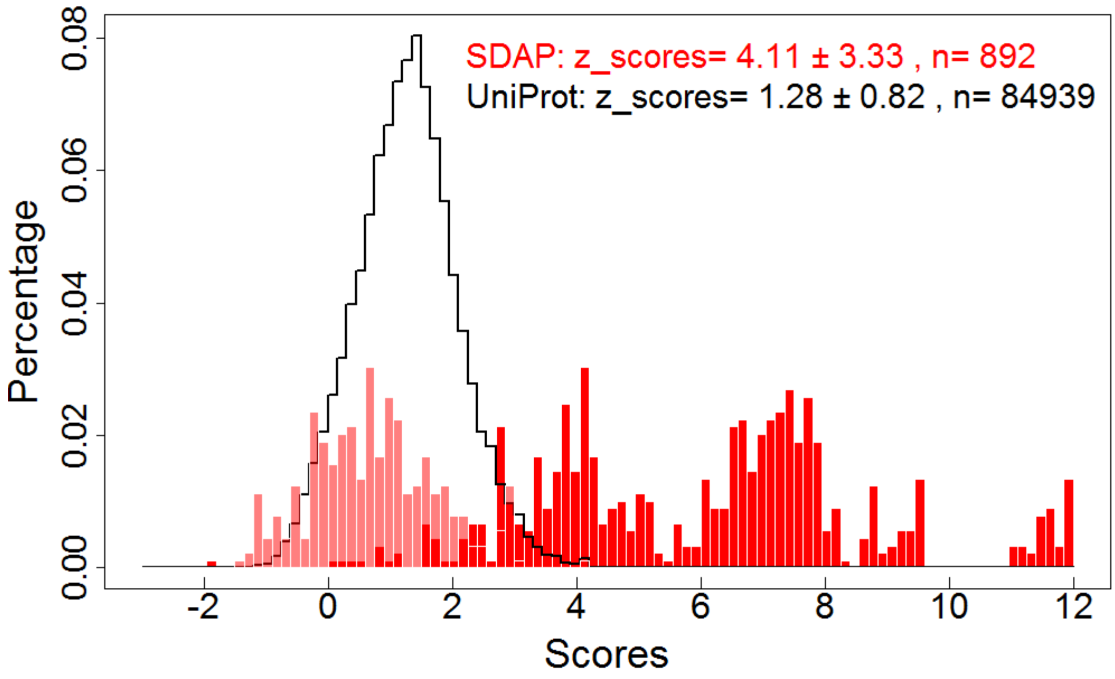


Figure 3.3 Distribution of z-scores for non-allergenic sequences and allergenic sequences.

Solid red: allergenic sequences from SDAP, 4.11 ±3.33, n=892; Black: non_allergenic sequences from uniprot, 1.28 ±0.82, n=84939; Light red: allergenic sequences not related to the 17 Pfam-A families.

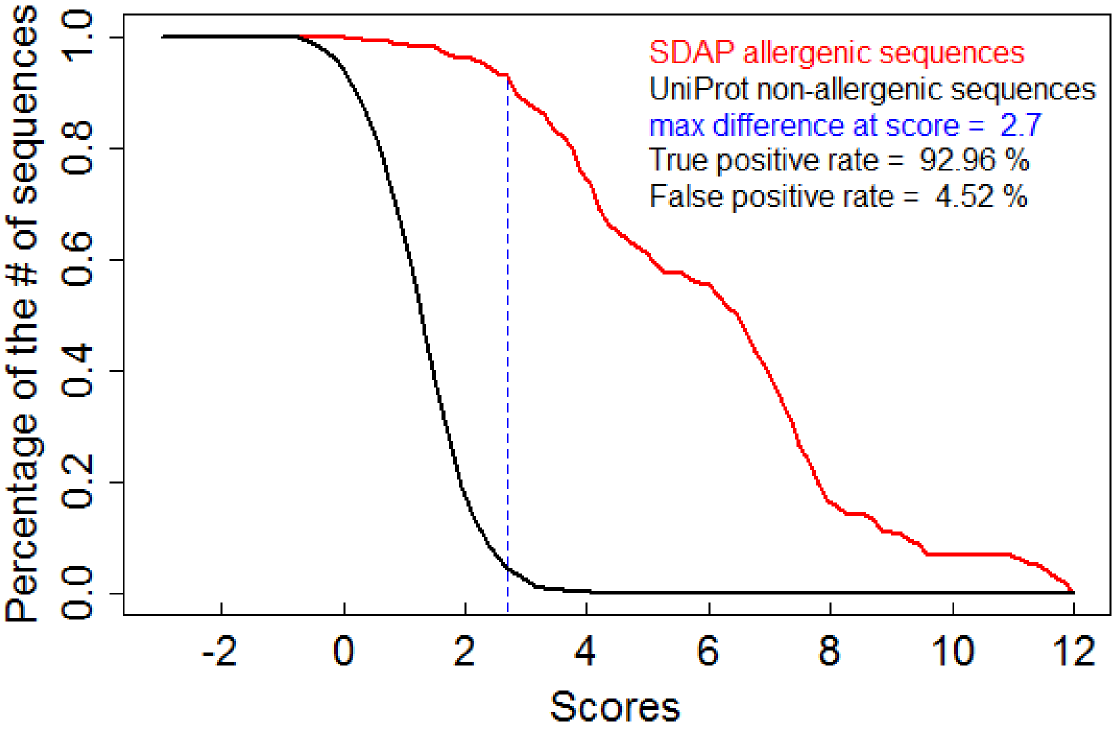


Figure 3.4 Cumulative percentage of non-allergenic sequences and allergenic sequences.

Red: allergenic sequences in the top 17 protein families from SDAP; Black: non_allergenic sequences from uniprot; Blue: maximum difference between the percentage of allergenic sequences and non-allergenic sequences at score = 2.7.

Motifs for protein families related to peanut allergens.

Peanuts are one of the major food sources for allergenic reactions and the symptoms are usually more severe than other food induced anaphylaxis (Al-Muhsen et al., 2003; Flinn and Hourihane, 2013). Approximately 1 % of children and 0.6 % of adults are affected by peanut allergy, while up to 3% of the population is sensitized to peanut allergens (Sicherer and Sampson, 2010). There are also reports indicating a trend for increase in reported peanut allergy (Grundy et al., 2002; Gupta et al., 2011; Rinaldi et al., 2012). Many studies have been done to identify the peanut allergens like Ara h 1-11 (Jiang et al., 2011) and a few experimental IgE binding epitopes have been reported (Shin et al., 1998; Shreffler et al., 2005; Stanley et al., 1997).

In our analysis, we have generated sequence motifs using the allergenic sequences from peanuts and their homologous sequences from Pfam-A protein families. These motifs are potentially specific to peanut allergens and are likely to match previously determined IgE epitopes. Thus, we used experimental methods of peptide microarrays to further validate the effect of our motifs in collaboration with Drs. C. H. Schein and S. J. Maleki (USDA, New Orleans). We also tried to identify overlaps between our allergen specific motifs and known experimental epitopes. For the test, we took the motifs from four Pfam-A families (PF00190**,** PF00234, PF00235 and PF00407) that contain major peanut allergens, Ara h 1-6 and 8. Short peptides according to these motifs were synthesized, and then we tested the binding affinity of these peptides to IgE in sera from patients who are known to be allergic to peanuts or tree nuts. The results show that many of our motifs overlap with previously reported IgE epitopes and they have high binding intensity to patients' IgE with maximum value of 1173 (Table 3.2). Meanwhile, there are also many motifs with low binding intensity, implying that these conserved motifs are probably responsible for other functions or structural stability of the protein. The experimental results indicate that our motifs are able to recognize known IgE binding epitopes, and could be used to provide suggestions for other potential IgE epitopes. Further improvements in our motif mining method are necessary to generate more refined motifs for the prediction of IgE epitopes.

Table 3.2 Binding intensity of motifs from major peanut allergens to IgE

The peptides were synthesized according one reference sequence of each Pfam-A family. Uniprot ID P43238 (Ara h 1) for PF00190. GeneBank ID 15418705 (Ara h 2) for PF00234. Uniprot ID Q9SQI9 (Ara h 5) for PF00235. GeneBank ID 37499626 (Ara h 8) for PF00407. The start and end number of the peptide in the SDAP sequences are shown at both sides of the motifs. Experimental epitopes are underlined.

* The experimental epitopes in PF00190 (Shin et al., 1998)

† The experimental epitopes in PF00234 (Stanley et al., 1997)

$ The experimental epitopes in PF00234 (Shreffler et al., 2005)

Pfam	Allergen motifs	Binding intensity
PF00190	23 RQFQNLQ 29	66
Cupin family	45 LPKHADA 51	51
	56 VIQQGQA 62	20
Ara h 1	76 NLDEGHA 82	81
Ara h 3	**116 QFEDFFPASSRDQ 128 (Epi#10)***	**527**
Ara h 4	**10 LSNNFGKLFEVKP 22 (Epi#15)**	**79**
	27 PQLQDLD 33	61
	49 PHFNSKA 55	56
	60 VVNKGTG 66	78
	106 RLKEGDV 112	93
	130 HLLGFGINAENNH 142 (Epi#17)	**175**
	166 FPGSGEQVE 174 (Epi#20)	**1173**
PF00234	**5 LMQKIQRD 12 (Epi#5)†**	**101**
Plant lipid transfer proteins	**47 CCNEL 51** $	**90**
Ara h 2, Ara h 6	60 CMCEAL 65	33
PF00235	2 WQTYVDNHLLC 12	77
Profilin family	26 GQDGGVWAQS 35	115
	38 FPQFKPEE 45	31
Ara h 5	56 PGSLAPTG 63	23
	71 YMVIQGE 77	67
	81 IIPGKKGPGGVT 92	78
	111 PGQCNM 116	42
	119 ERLGDY 124	23
PF00407	1 MGVFTFEDE 9	19
Pathogenesis-related	25 DADSITPK 32	42
protein Bet v I family	45 GNGGPGT 51	56
	69 KVESID 74	20
Ara h 8	77 NYAYNYS 83	24
	100 ETKLVEGPNGGS 111	NA
	119 YHTKG 123	79
	129 EEELK 133	50
	145 AIEGYVLA 152	23

Scores of sequences from human microbiome

There are numerous microorganisms living in the human body that play fundamental roles in human health and diseases. They include microbes like eukaryotes, archaea, bacteria and fungi whose number is estimated to be higher than the number of human cells by an order of magnitude (Relman and Falkow, 2001; Savage, 1977). Through international collaboration and extensive effort of high throughput sequencing, the database of human microbiome (Gevers et al., 2012; Group et al., 2009; Human Microbiome Project, 2012) collects sequences of these microbial communities found at multiple body sites like airways, skin, oral cavity, gastrointestinal tract and vagina of healthy human, aiming to provide a baseline view of the healthy human microbiome. Most of these microorganisms are involved in normal biological processes in human body. Considering that they are able to mutually exist in our body without causing any immune response or allergic reaction, they could be used as sources of non-allergenic sequences. I have found many homologues of allergenic proteins in the database of human microbiome and present here the result on the family of pectate lyases that contain both allergenic and non-allergenic proteins.

Pectate lyases are enzymes involved in degradation of plant cell wall (Marin-Rodriguez et al., 2002). They are mainly found in plant pathogens and many plant tissues including pollen and ripening fruits (Wing et al., 1990). Bacteria in the digestive tract of animals who feed on plant also produce pectate lyases (Hugouvieux-Cotte-Pattat, 2014). Multiple pollen allergens are found in the family of pectate lyases, including Amb a 1, Cup a 1, Jun a 1 et.al (Aceituno et al., 2000; Griffith et al., 1991; Midoro-Horiuti et al., 1999).

Meanwhile, bacteria in human microbiome secrete pectate lyases to digest plant materials in guts. Thus, the homologous sequences in the family of pectate lyases are composed of both allergenic and non-allergenic sequences, making it a good test dataset for our method.

We blasted the sequences of human microbiome database with mountain cedar allergen Jun a 1, and identified 18 homologous sequences with E-value below the magnitude of 10^{-2} as non-allergenic sequences (Table 3.3). The sequences in the pectate lyase family (PF00544) were obtained from SDAP as allergenic sequences. All sequences of PF00544 were downloaded from Pfam as background. Scores were calculated for all three sets of sequences against our allergen-specific motifs. The distribution indicates a clear separation between allergenic sequences in PF00544 (red) and their homologues in human microbiome (green) (Figure 3.5). On the other hand, the two criteria of FAO/WHO correctly detected all allergenic sequences, but they also reported 15 positives out of 18 non-allergenic homologues with a false positive rate of 83%. This demonstrates the ability of our method to minimize false positive rates while recognizing all allergenic sequences. Besides the known allergens, there might be other potential allergens in this protein family and it is of great interest for researchers to detect them. The scores for all the sequences in PF00544 show a bimodal distribution with one group clustered close to allergenic sequences at high scores and another group close to the non-allergenic side (black). The clusters of two groups at different scores against allergen-specific motifs may imply separation between potential allergenic sequence and non-allergenic ones.

Table 3.3 Homologues of pectate lyase in Human Microbiome

Seq_ID	Protein name	Organism	E-Value
HMPREF0004_0963	conserved hypothetical protein	Achromobacter piechaudii ATCC 43553	2.00E-13
HMPREF0005_04335	hypothetical protein	Achromobacter xylosoxidans C54	1.00E-12
HMPREF9321_1726	putative pectate lyase	Veillonella atypica ACS+049+V+Sch6	7.00E-11
HMPREF0908_0299	pectate lyase	Selenomonas flueggei ATCC 43531	2.00E-10
HMPREF1019_01313	hypothetical protein	Campylobacter sp. 10_1_50	5.00E-10
HMPREF9166_1475	pectate lyase	Selenomonas sp. oral taxon 149 str. 67H29BP	7.00E-10
HMPREF1012_01654	pel protein	Bacillus sp. BT1B_CT2	1.00E-08
HMPREF7545_0107	pectate lyase	Selenomonas noxia ATCC 43541	2.00E-08
HMPREF1012_03545	pectate lyase	Bacillus sp. BT1B_CT2	3.00E-06
RUM_11790	Pectate lyase	Ruminococcus champanellensis 18P13	7.00E-06
HMPREF9412_2007	pectate lyase	Paenibacillus sp. HGF5	2.00E-05
CUW_2559	pectate lyase	Turicibacter sanguinis PC909	1.00E-04
BACPEC_00859	pectate lyase	Clostridiales	2.00E-04
HMPREF9402_0417	pectate lyase	Turicibacter sp. HGF1	2.00E-04
HMPREF6485_0196	CHU large protein candidate pectate lyase	Prevotella buccae ATCC 33574	2.00E-03
HMPREF0649_01089	CHU large protein candidate pectate lyase	Prevotella buccae D17	2.00E-03
HMPREF1146_2484	pectate lyase	Prevotella sp. MSX73	2.00E-03
PREVCOP_04892	pectate lyase related protein, secreted	Prevotella copri DSM 18205	2.00E-03

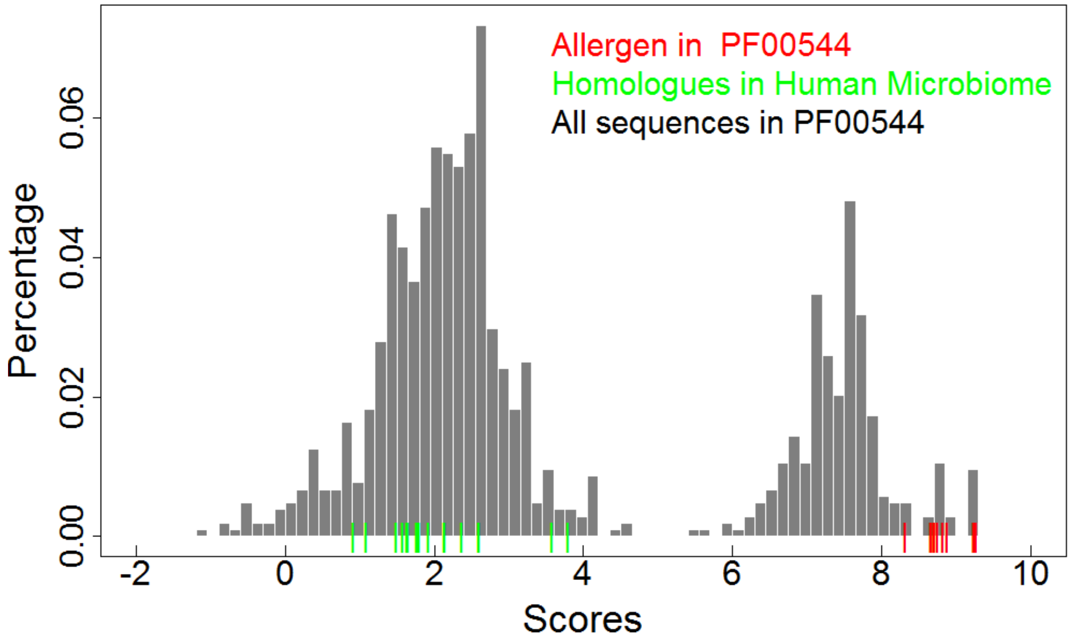


Figure 3.5 z-scores for allergenic sequences in PF00544 (Pectate Lyase) and their homologues in Human Microbiome.

Red: z-scores for allergenic sequences from SDAP, n=17; Green: z-scores for homologues in Human Microbiome, n=18; Black: distribution of z-scores for all sequences from PF00544, 3.55 ± 2.56, n=1037

Correlation between the cross-reaction results of allergen chip and their allergenicity scores

Cross-reaction is a well-established fact that occurs when an individual allergic to certain allergen is exposed to other allergens that share similar sequence or structure with the original allergen. Detection of cross-reactivity between allergens is of great interest because it could extend across a wide range of species (Midoro-Horiuti et al., 2003) and may lead to severe symptoms when individuals unknowingly consume foods containing cross-reactive allergens (Jenkins et al., 2005). One well-studied example is the birch pollen allergen Bet v 1, which is cross-reactive with several fruits and nuts allergens like Pru av 1 from cherry, Mal d 1 from apple, Gly m 4 from soybean, and Ara h 8 from peanut (Bolhaar et al., 2005; Egger et al., 2006; Mittag et al., 2004; Mittag et al., 2006; Mittag et al., 2005). A study by Pfiffner et.al. tested the cross-reactivity between Bet v 1 and several other allergens by measuring the binding of patients' serum samples with allergen chips coated with different allergenic proteins (figure 5c in (Pfiffner et al., 2012)). They reported the table with percentages of patients that are positive against allergens in rows who are also positive against allergens in columns. Their results revealed a hierarchy in the degree of cross-reactivity among allergens, where 6 allergens, Bet v 1, Cor a 1.0401, Cor a 1.0101, Pru p 1, Aln g 1 and Mal d 1 share a greater portion of patients who are positive to them than the other allergens.

To verify if we could identify similar hierarchy of cross-reactivity in the allergenicity scores of these allergens, we collected 10 allergenic sequences corresponding to these allergens and calculated the z-scores of them against allergen-specific motifs from their Pfam-A family PF00407 (Table 3.4). The scores for the top 6 allergens are close to each other while the other 4 allergens scored below 6. To directly compare our z-scores with experimental results, we attempted to transform the original data composed of percentages to distance matrixes representing the extent of cross-reactivity (Table 3.5). For each pair of allergens a and b in the original table, P_{ab} denotes the percentage of patients that are positive against allergen a who are also positive against allergen b while P_{ba} is the other way around. P_{ab} does not necessarily equal to P_{ba} because the extent of cross-reactivity varies among allergens. The distance D between the pair of allergens a and b was calculated by geometric mean,

$$D = -log\sqrt{P_{ab} \cdot P_{ba}}$$

Therefore, when there are higher percentages of patients positive to the pair of allergens, the distance is smaller. We further visualized the distance matrix by 2 dimensional multidimensional scaling (MDS) to show a “map” of the allergens corresponding to experimental determined cross-reactivities (Figure 3.6). The data points on the map were color coded according their allergenicity scores with red for high scores and blue for low scores. The plot clearly shows the cluster of top 6 allergens on the top left corner and they are mostly with higher scores in red color, suggesting an agreement on estimating cross-reactivity between our scores and the experimental test.

Table 3.4 Scores of the 10 allergens in PF00407

Allergen	*uniprot*	*pfam*	*z-scores*
Bet v 1	P45431	BEV1B_BETPN	6.94
Cor a 1.0401	Q9SWR4	Q9SWR4_CORAV	7.12
Cor a 1.0101	Q39454	Q39454_CORAV	7.24
Pru p 1	Q2I6V8	Q2I6V8_PRUPE	7.12
Aln g 1	P38948	MPAG1_ALNGL	7.01
Mal d 1	P43211	MAL11_MALDO	6.13
Ara h 8	B0YIU5	B0YIU5_ARAHY	5.35
Gly m 4	P26987	SAM22_SOYBN	5.72
Api g 1	P49372	ALL1_APIGR	5.26
Dau c 1	O04298	DAU1_DAUCA	5.09

Table 3.5 Distance matrix generated from the cross-reaction results

	Bet v 1	Cor a 1.0401	Cor a 1.0101	Pru p 1	Aln g 1	Mal d 1	Ara h 8	Gly m 4	Api g 1	Dau c 1
Bet v 1	0	0.416	0.216	0.27	0.239	0.223	0.466	0.783	0.836	1.488
Cor a 1.0401	0.416	0	0.434	0.384	0.414	0.401	0.604	0.828	0.933	1.512
Cor a 1.0101	0.216	0.434	0	0.295	0.255	0.295	0.444	0.663	0.84	1.387
Pru p 1	0.27	0.384	0.295	0	0.236	0.211	0.361	0.647	0.737	1.285
Aln g 1	0.239	0.414	0.255	0.236	0	0.249	0.376	0.647	0.707	1.368
Mal d 1	0.223	0.401	0.295	0.211	0.249	0	0.378	0.647	0.719	1.285
Ara h 8	0.466	0.604	0.444	0.361	0.376	0.378	0	0.461	0.589	1.171
Gly m 4	0.783	0.828	0.663	0.647	0.647	0.647	0.461	0	0.683	1.117
Api g 1	0.836	0.933	0.84	0.737	0.707	0.719	0.589	0.683	0	0.606
Dau c 1	1.488	1.512	1.387	1.285	1.368	1.285	1.171	1.117	0.606	0

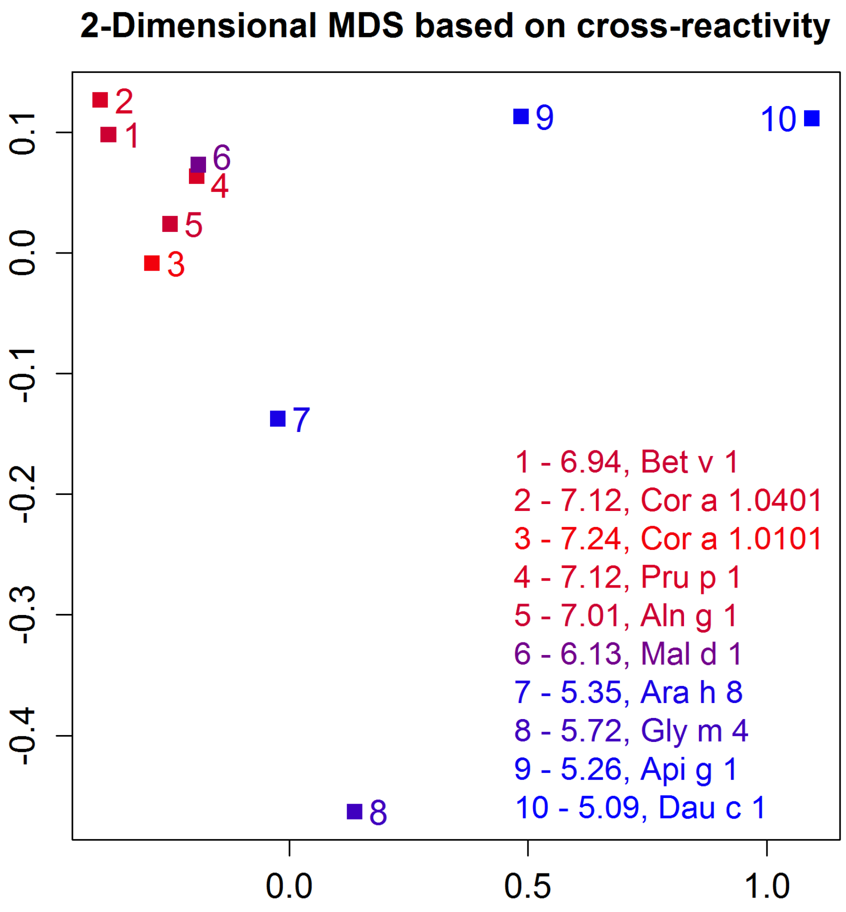


Figure 3.6 2-dimensional MDS of distance matrix of cross-reactivity.

The colors of data points represent allergenicity scores where red is high score and blue is low score. The score values and allergen types are listed in the legend.

3.4 Discussion

Allergic reactions are common diseases that could cause mild to lethal symptoms and affect millions of people in the US. Allergens could be found in various sources in the environment of our daily lives, like pollen, insects, drug and food et.al. Nowadays, due to the introduction of new materials or genetically modified products into the environment and our food sources, the risk of introducing new sources of allergens also increases. Thus, a lot of effort has been spent on establishing methods and guidelines to evaluate allergenicity of proteins and identify potential allergens.

The criteria suggested by FAO/WHO have been implemented in servers like SDAP (Ivanciuc et al., 2003) and Allermatch (Fiers et al., 2004). However, it has been shown that these criteria tend to over-predict allergenicity and modification of the cutoff values was suggested (Goodman, 2008; Ladics et al., 2007). The drawback of FAO/WHO guidelines implies that assessment of allergenicity is more complicated than simply comparing sequence identity or finding identical peptides. Various more sophisticated bioinformatics tools have been developed to achieve the prediction of allergenicity and most of them outperform the FAO/WHO guidelines in terms of false positive rates. These new methods could be roughly categorized into two types of approaches.

The first type is the traditional motif-based method. Stadler et.al. generated 52 allergen-relevant motifs from allergenic sequences using MEME and searched for matches of the motifs in query sequences (Pfiffner et al., 2012; Stadler and Stadler, 2003). WebAllergen

(Li et al., 2004; Riaz et al., 2005) clustered the allergenic sequences first and then identified motifs using wavelet transform and Hidden Markov Model (HMM) profiles. However, the motifs used by both methods are long peptides with length of ~50 (MEME) or >30 (WebAllergen) amino acids, which rather represent the characters of the protein domains or families instead of allergenic features (Schein et al., 2007).

The second type of method is based on learning and classification algorithms, especially Support Vector Machine (SVM). Many applications of this method have been developed recently by constructing different feature vectors as listed below (Table 3.6). Some of these methods mainly rely on decomposing sequences and generating vectors containing frequencies of amino acids and short peptides (AlgPred (Saha and Raghava, 2006), Allerdictor (Dang and Lawrence, 2014)) or similarities to allergen-relevant peptides of different length (EVALLER (Martinez Barrio et al., 2007; Soeria-Atmadja et al., 2006), SORTALLER (Zhang et al., 2012), AllerHunter (Muh et al., 2009)). The accuracy of these methods highly depend on sequence similarity to allergenic proteins and they may perform poorly on distinguishing allergenic and non-allergenic sequences that are homologous (Schein et al., 2007). APPEL (Cui et al., 2007) and PREAL (Wang et al., 2013) noticed the potential role of physical-chemical properties in allergenicity and included global properties for proteins like hydrophobicity, van der Waals volume, polarity, molecular weight, length and subcellular locations in their feature spaces for SVM. However, the overall physical-chemical properties of proteins may imply their general characteristics or functions and are probably not specific to their allergenicity.

Another important factor that needs consideration in predicting allergenicity is the cross-reactivity between allergens. Allergenic proteins are from various sources and with very different functions or structures. As shown by the allergens from PF00407 (Pfiffner et al., 2012), most cross-reactive allergens are actually from the same protein families, suggesting there are homologies between allergens. Thus, grouping allergenic sequences by protein families facilitates the identification of homologies among related allergens.

Our method to evaluate allergenicity is based on Physical-Chemical Property (PCP) motifs generated specifically for different families of allergenic protein. We classified the allergenic sequences according to their Pfam-A families to ensure the motifs are more accurate and specific to different allergens. The motifs are based on conservation of properties instead of amino acids, which enables detection of more subtle homologies. For allergens in each family, there are multiple motifs of short length (5-15 amino acids) serving as fingerprints to reduce the rate of false positives. Therefore, our method has better ability in distinguishing allergenic and non-allergenic sequences even when they are homologous.

Table 3.6 Summary of SVM-based allergenicity prediction methods.

Name	*Input vectors for SVM*
AlgPred (Saha and Raghava, 2006)	amino acid and dipeptide composition of proteins
Allerdictor (Dang and Lawrence, 2014)	frequency of k-mer peptide with best performance at k=6
EVALLER (Martinez Barrio et al., 2007; Soeria-Atmadja et al., 2006)	alignment scores against Filtered Length-adjusted Allergen Peptides (FLAPs) whose lengths are variable with minimum of 22
SORTALLER (Zhang et al., 2012)	normalized BLAST E-value between proteins and a set of Allergen Family Featured Peptides (AFFP) with lengths ranging from 20 to > 200 amino acids
AllerHunter (Muh et al., 2009)	sequences similarity to a list of known allergen and putative non-allergen sequences
APPEL (Cui et al., 2007)	vectors of 41 elements combining the 21 elements of global physicochemical properties and the 20 elements of amino acid composition
PREAL (Wang et al., 2013)	vectors of 128 elements including similar elements of physicochemical properties and amino acid composition to APPEL, 1 element for molecular weight, 1 element for length of protein and 22 elements of Boolean values for protein's subcellular locations

3.5 Conclusions and future studies

We have developed a novel method to evaluate potential allergenicity of query sequences using allergen-specific motifs for major allergens in the top 17 Pfam-A families. The scoring method is able to distinguish non-allergenic sequences and allergenic sequences in these families. The motifs from several major peanut allergens overlap with experimental IgE epitopes and many of the peptides synthesized according to the motifs show high binding efficiency to patient’s serum, indicating their importance in IgE binding and allergic reactions. The distributions of allergenicity scores for allergens in PF00544 and their homologous non-allergenic sequences from human microbiome show clear separation. The scores could also demonstrate the hierarchy of cross-reactivity between allergens in PF00407 as implied by experiments.

The scoring method using the motif profiles from known allergenic sequences seems to be a promising approach to evaluate allergenicity; however, due to the limitation of data, there are not enough sequences in many protein families to generate meaningful motifs, thus, the method is not working efficiently for allergenic sequences without characteristic motifs. In the future, we aim to establish motif profiles for most protein families and improve the accuracy of the scores. In addition to linear motifs, we also plan to identify conformational motifs with the collection of experimental structures and models in SDAP. Finally, we aim to provide comprehensive tools to assist the public, physicians, government regulators and companies in evaluating the risk of allergenicity and minimizing the prevalence of allergic diseases.

CHAPTER 4: COMPUTATIONAL DESIGN OF VACCINES AGAINST MULTIPLE STRAINS OF ENCEPHALITIC ALPHAVIRUSES

Outbreaks of encephalitic alphaviruses have caused severe epidemics in Asia and more recently in Europe, and several viral strains are endemic in the US (Weaver et al., 2012); however, no vaccine is currently licensed for mass immunization (Metz and Pijlman, 2011). The attenuated vaccine strain, TC-83, of Venezuelan equine encephalitic virus (VEEV), used for immunization of horses and lab personnel has significant side effects and environmental risks (Paessler and Weaver, 2009). The design of a multivalent vaccine against several species remains a challenge because of the high sequence variability among different viral species. We illustrate here our new computational approach for a design strategy for common protection against different strains of VEEV and eastern equine encephalitic alphaviruses (EEEV) species. The computational approach is based on sequences of VEEV and EEEV from a Next Generation Sequence effort, a detailed analysis of the variability and conservation of physical-chemical properties (PCP) of residues in the E2 envelope proteins, structural information from Cryo-EM and X-ray studies of CHIKV, VEEV and 3D models of EEEV, and prediction of conformational epitopes. Specific constructs of the mosaic E2 protein were suggested for future experimental tests in an animal challenge model against VEEV and EEEV.

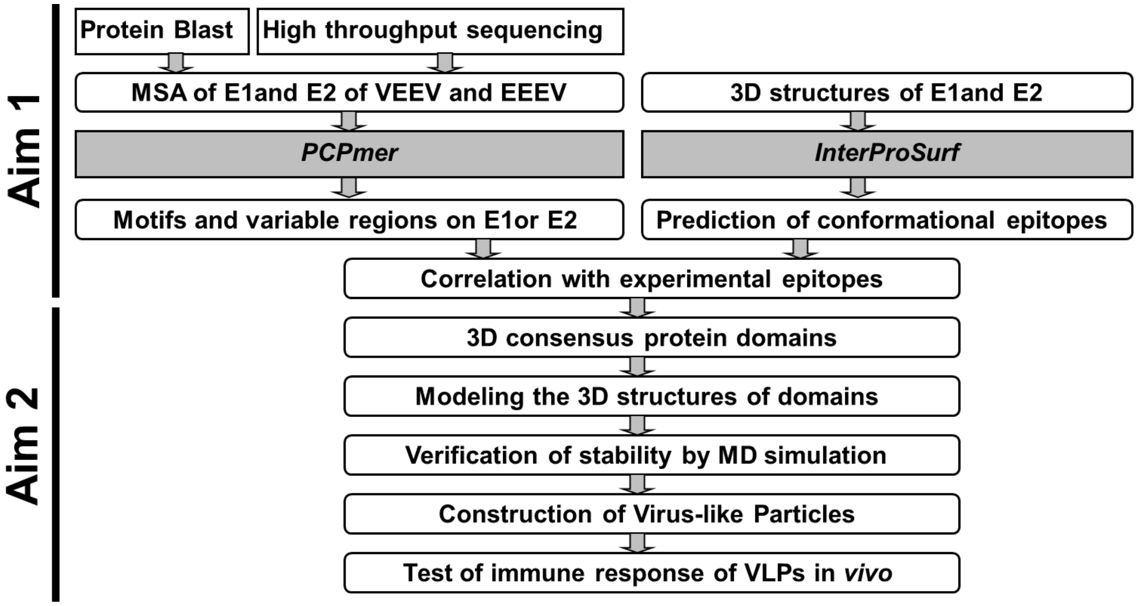

Aim 1
Aim 2
Protein Blast
High throughput sequencing
MSA of E1and E2 of VEEV and EEEV
3D structures of E1and E2
PCPmer
InterProSurf
Motifs and variable regions on E1or E2
Prediction of conformational epitopes
Correlation with experimental epitopes
3D consensus protein domains
Modeling the 3D structures of domains
Verification of stability by MD simulation
Construction of Virus-like Particles
Test of immune response of VLPs in vivo

Figure 4.1 Overview of the strategy to design multivalent vaccine

4.1 Introduction

Alphaviruses

Alphaviruses are positive sense, single-stranded RNA viruses that include 29 different species such as Semliki Forest, Chikungunya (CHIKV), Barmah Forest, Western equine encephalitis (WEEV), Eastern equine encephalitis (EEEV) and Venezuelan equine encephalitis viruses (VEEV) (Figure 4.2) (Weaver et al., 2012). They are mainly transmitted by mosquitoes and infect a wide variety of vertebrates like humans, horses, rodents, birds and fish. Alphaviruses are significant human and veterinary pathogens which are responsible for several medically important emerging diseases in human and domestic animals. Massive outbreaks of various species have been documented in multiple continents that involved millions of people and equines, causing huge economic losses and great burdens on healthcare. For example, CHIKV is considered to be one of the most important human pathogen among alphaviruses due to its high morbidity rate and potential mortality. The recent global outbreak started as an epidemic in Kenya, Africa in 2004 and the viruses spread to nearby islands and India by late 2005, resulting in explosive epidemics (Tsetsarkin et al., 2011). Then, the viruses were further transmitted to Southeast Asia in 2006 (Noridah, 2007), Italy in 2007 (Rezza et al., 2007) and France in 2010 (Grandadam et al., 2011) by infected travelers from epidemic regions. There were also a few infected travelers identified in the United States (Lanciotti et al., 2007), and fortunately no massive epidemic was reported. This outbreak is still ongoing with recent cases reported in the Caribbean islands and may further spread to the American mainland (Leparc-Goffart et al., 2014). So far, this epidemic of CHIKV has spread over large regions and affected hundreds-of-thousands of people with mild cases of malaise, fever, headache and severe cases of chronic joint pains and neurological sequelae that last for years (Gerardin et al., 2011; Powers and Logue, 2007; Tsetsarkin et al., 2011). Fatal cases of CHIKV infections were also reported especially in underdeveloped regions (Mavalankar et al., 2008).

Encephalitic Alphaviruses, VEEV and EEEV

The species that strongly affect Central, South America and the U.S. are WEEV, EEEV and VEEV, which cause severe symptoms of encephalitis. These viruses were first found back in 1930s from infected horses in California (Meyer et al., 1931), Virginia and New Jersey (Giltner, 1933; TenBroeck, 1933), and from an infected child in Caracas, Venezuela (Beck and Wyckoff, 1938), and therefore named according to their regions of isolations. The symptoms of these viruses are similar in infected patients and the morbidity and mortality rates are much higher in younger children. After a few days of incubation period, the symptoms start with mild fever, headache or nausea at early phases, and some infected patients further develop severe edema, congestion or hemorrhages in the brain, gastrointestinal tract or lungs that lead to death or neurological sequelae (de la Monte et al., 1985). The fatality rate of WEEV is 3-7% while 15-30% of the survivors of encephalitis are estimated to suffer from neurological sequelae (Zacks and Paessler, 2010). EEEV is considered to be the most virulent encephalitic alphaviruses with a fatality rate over 50% (Zacks and Paessler, 2010), although infection of humans with EEEV is less frequent than with WEEV according to the number of cases reported by the U.S. Centers for Disease Control. VEEV includes multiple antigenic subtypes (IAB, IC, ID, IE, IF) with IAB and IC being responsible for most epidemics/epizootics in human and equines. In human cases, about 5–15% of patients, especially children, develop neurological symptoms with an overall case-fatality rate of about 0.5% (Weaver et al., 2012). Between 1930s and early 1970s, periodic outbreaks of VEEV occurred in South America every ~10 years and each outbreak lasted months to years (Lord, 1974; Weaver et al., 2004). The most recent outbreak of subtype IC in Venezuela and Colombia in 1995 involved an estimated 75 000 to 100 000 people and 3000 fatal cases (Weaver et al., 1996).

In addition to the threat of naturally emerging encephalitic alphaviruses, EEEV and VEEV were also researched to develop biological weapons during the cold war (Bronze et al., 2002; Hawley and Eitzen, 2001) because that both viruses are extremely infectious via the aerosol route and the resulting symptoms cause either death or neurologic sequelae that incapacitate patients (Reed et al., 2007; Reed et al., 2004). Even though great progress has been made in understanding the pathogenesis of encephalitic alphaviruses, the current treatments are mainly to relieve symptoms and there are no effective antiviral treatments to clear viruses especially when viruses reach the brain (Weaver et al., 2012; Zacks and Paessler, 2010). Therefore, there is a pressing need for vaccines against encephalitic alphaviruses, not only due to their potential to cause massive outbreaks, but also the importance of biodefense during warfare or terrorism attacks.

The attempts to develop vaccines started since the isolation of encephalitic alphaviruses in 1930s. After the outbreak in 1938, the first inactivated VEEV vaccines were prepared from animal tissues infected with isolated viruses and inactivated by formalin (Randall et al., 1949). However, the isolation of residual live viruses from these vaccines indicated high risks of incomplete inactivation (Sutton and Brooke, 1954). Later, an attenuated VEEV vaccines, TC-83, was developed by serial passaging VEEV subtype IAB strain, Trinidad Donkey, for 83 times in guinea pig heart cells (Berge TO, 1961). TC-83 was extensively used for limiting the spread of viruses in equids during the outbreak between 1969 and 1971 in Central America and Texas (Walton et al., 1972). However, further tests of TC-83 on humans induced seroconversions on 80% of humans while causing symptoms like fever and headache to ~25% of the volunteers (Paessler and Weaver, 2009). Currently, attenuated TC-83 was restricted to the vaccination of laboratory researchers and military personnel who are at risk of exposure to viruses, and further vaccination of C-84, a formalin-inactivated version of TC-83, is needed for people with no signs of neutralizing antibodies after initial vaccination (Pittman et al., 1996). Many efforts have been spent on improving TC-83 or develop vaccines using other methods

(Dupuy et al., 2009; Paessler et al., 2003; Pratt et al., 2003; Volkova et al., 2008). However, most of these vaccines only protect against one or two subtypes, it remains a challenging task to develop safe, immunogenic and efficacious vaccines against multiple encephalitic alphaviruses.

In this study, I aim to design multivalent vaccines against VEEV and EEEV. The sequences of the envelope proteins were analyzed by PCPMer (Mathura et al., 2003) to identify conserved and variable regions. Conformational epitopes that are critical in the binding of antibodies were predicted by our tool, InterProSurf (Negi et al., 2007) with the experimental structures and homology models. The predicted epitopes and experimental epitopes were mapped on the surface of envelope proteins. Finally, hybrid proteins were designed to represent the immune features of both VEEV and EEEV, and we expect to use them as vaccines to induce neutralizing antibodies against both encephalitic alphaviruses.

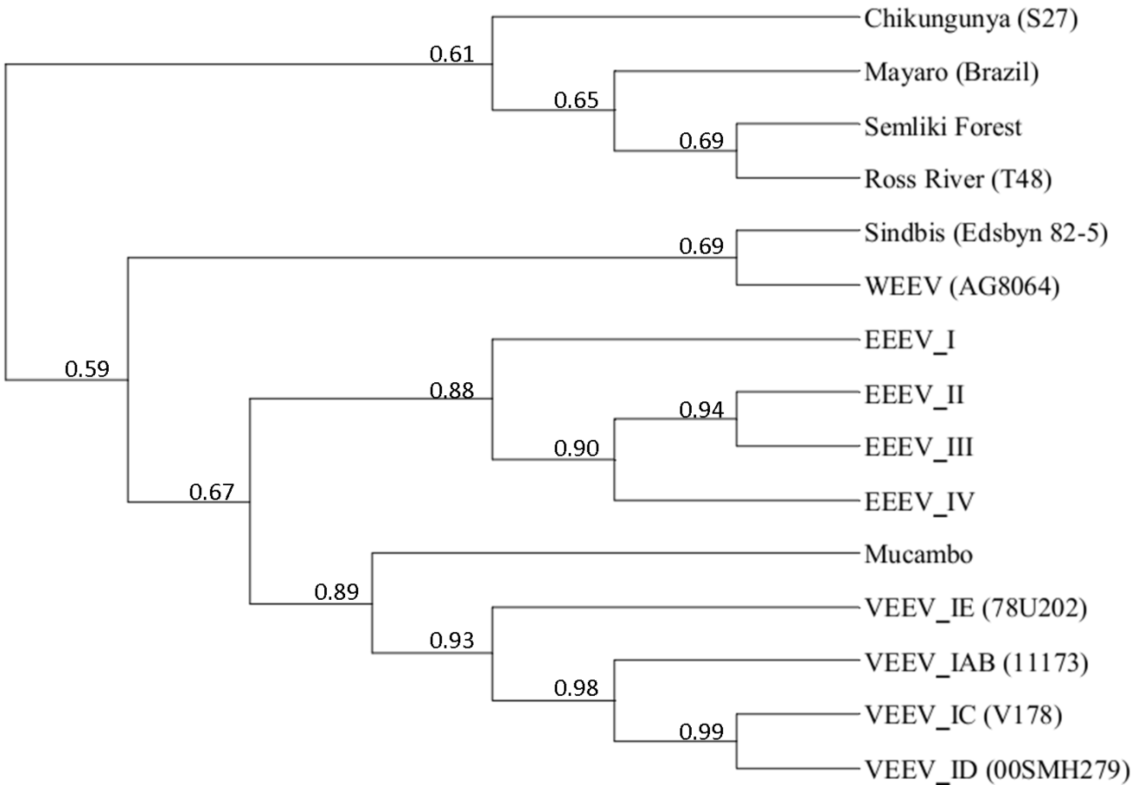


Figure 4.2 Phylogenetic tree of major species of alphaviruses.

The names of the strains are shown in the brackets. The sequences identities of envelope proteins E2 are shown at the branch.

Envelope proteins of alphaviruses

The 3D structures of alphaviruses provide essential information for understanding the mechanism of virus entry, replication and assembling, also shed light on the development of vaccines. There has been substantial progress in solving their structures by X-ray crystallography and cryo-electron microscopy (cryo-EM). Several crystal structures of the structural proteins have been obtained, such as the envelope protein E1 of Semliki Forest virus (Roussel et al., 2006) and the E3/E2/E1 complex of Chikungunya virus

(Voss et al., 2010). Additional cryo-EM structures of the complete virus particles for VEEV (Zhang et al., 2011), Barmah forest virus (Kostyuchenko et al., 2011), Sindbis virus (Mukhopadhyay et al., 2006; Tang et al., 2011) and Chikungunya virus (Sun et al., 2013) helped to understand the overall organization of these envelope proteins on the surface.

The complete genome of alphaviruses ranges between 11.4–11.8 kB in length, encoding non-structural proteins (nsP1-nsP4) that are responsible for RNA transcription and replication, and structural proteins forming the capsid and glycoprotein shell to protect the genome (Vaney et al., 2013). The structural proteins include capsid protein C and the envelope glycoproteins E3, E2 and E1. The capsid protein C forms an icosahedral nucleocapsid containing the RNA genome. The glycoprotein envelope is a spherical shell with 80 spikes on the surface and each spike is a trimer composed of three E1/E2 hetero-dimers. The spikes are further anchored to capsid proteins through the helical tails of glycoproteins E1 and E2 across bilayer of lipids. E1 contains three beta-barrel domains (I-III) with a fusion loop at the distal end of domain II. E2 is composed of three immunoglobulin (Ig) superfamily domains (A, B, C) with domains A and B forming a groove that covers the fusion loop of E1. E3 is a part of the precursor of E2, p62, and is often cleaved from the mature virus. It is suggested to be required for the formation of p62/E1 heterodimer in the Endoplasmic Reticulum of the infected cells during virus assembling (Lobigs et al., 1990; Salminen et al., 1992). Envelope proteins E1 and E2 play critical roles in the entry of virus into host cells, in which E2 protein binds to certain receptors on cell membranes, and then the low pH condition in endosome triggers conformational changes that dissociate E1/E2 heterodimers and form homotrimers of E1.

Then the exposed hydrophobic fusion loop of E1 protein inserts into the endosome membrane and the capsid and genomic RNA are released into the cytoplasm for replication (Li et al., 2010).

Among the envelope proteins, E2 is of particular interest for vaccine development because domains A and B of E2 are located on the top of the spike where they are easily accessible by receptors or antibodies. Thus, vaccines can be designed to either bind the antibody epitopes or block the receptor binding sites (Vaney et al., 2013). For example, Roehrig et.al analyzed the escape mutants to neutralizing monoclonal antibodies (MAbs) and identified the murine MAbs binding epitopes in the region of aa182-207 (Roehrig and Mathews, 1985) and human MAbs binding epitopes between aa115-119 (Hunt et al., 2010) of VEEV E2 protein. Similar studies on Sindbis virus (Strauss et al., 1991), Ross River virus (Kerr et al., 1992) and Chikungunya virus (Kam et al., 2014; Kam et al., 2012; Sun et al., 2013) also indicate MAbs binding sites on E2 protein. Therefore, I will focus on E2 proteins for the design of multivalent vaccines against encephalitic alphaviruses.

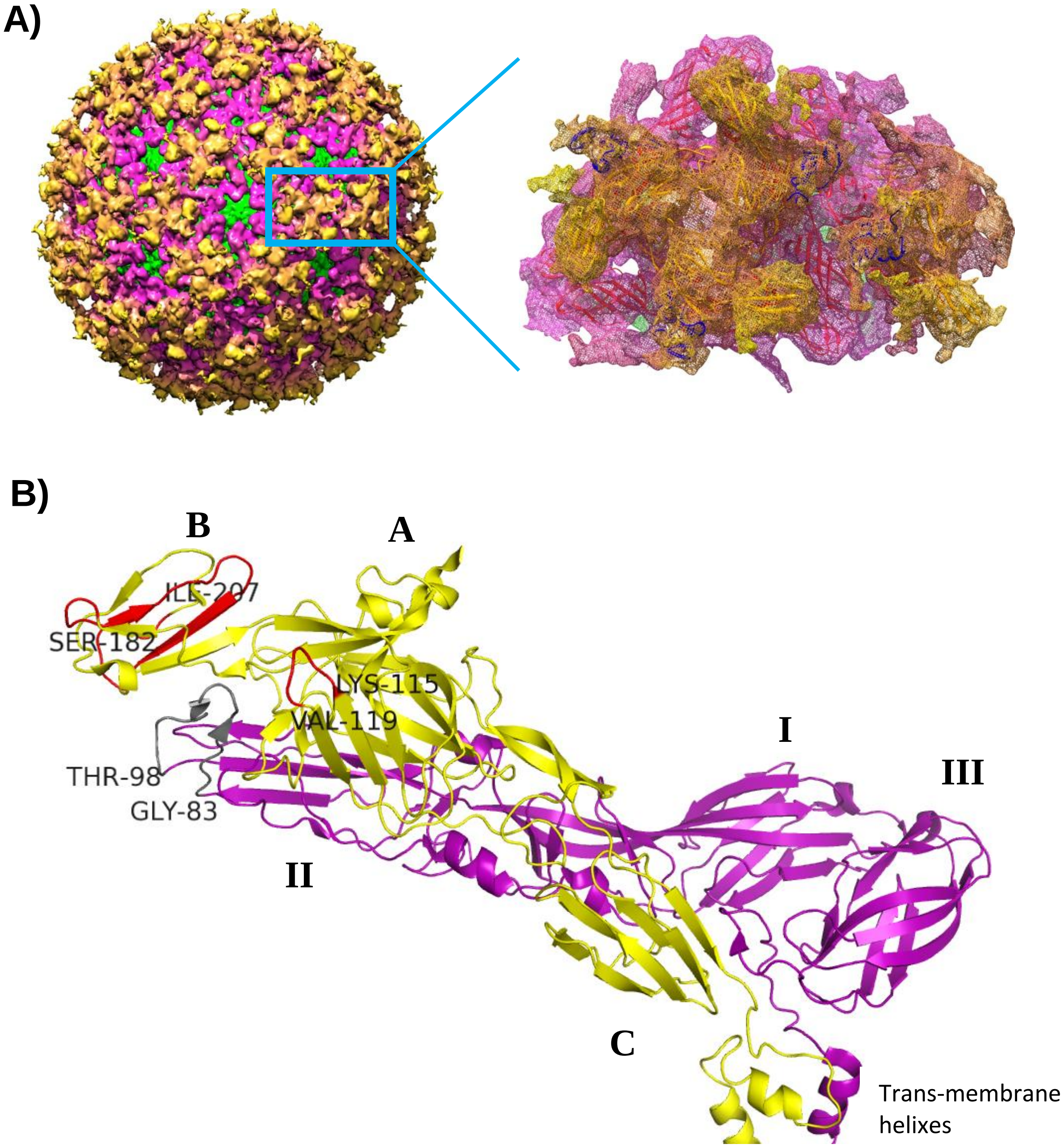


Figure 4.3 Envelope glycoproteins of VEEV subtype II, Everglades.

A). The structure was determined to 10 Å resolution by cryo-EM density map in collaboration with Dr. Michael B. Sherman (Sherman et al., 2013). The envelope glycoproteins form 80 trimer spikes, each spike consisting of 3 glycoprotein E1/E2 hetero-dimers (magenta and yellow). E3 is a component of E2 which is cleaved in the mature VEEV (blue).

B). The ectodomains of E1/E2 hetero-dimer (side view). The virus attaches to host cell receptors though the E2 (yellow) glycoprotein and the E1 (magenta) protein includes a fusion peptide (aa 83-98, grey) that mediates entry of nucleocapsids into the cytoplasm from endosomes. The three domains of E1 (I, II, III) and E2 (A, B, C) are labeled. Previously determined epitopes that bind murine monoclonal antibodies (mMAbs, aa 182-207) and human Mabs (aa 115-119) were mapped on E2 (red).

4.2 Methods

Phylogenetic tree

The phylogenetic tree was generated from 15 representative sequences of the envelope protein E2 to show the evolutional relationship between major alphaviruses species. The representative sequence of VEEV and EEEV were taken from the sequencing results of Dr. Scott Weaver's group and the other sequences were obtained from UniProt database. The multiple sequences alignment and distances were calculated by *ClustalW2* (Larkin et al., 2007) using Neighbor-joining method. The phylogenetic tree was then visualized by *SeaView* (Gouy et al., 2010). The average sequence identities were calculated for each branch of the tree.

PCP distance and similarity

The multiple sequence alignments of VEEV and EEEV were analyzed to identify conserved and variable regions. For each column of the alignments, the variability is defined by the average distance of physical-chemical properties of the amino acids in that column, D

$$D = \frac{2}{N(N-1)} \sum_{i \le j}^{N} \sqrt{\sum_{p=E1}^{E5} (V_i^p - V_j^p)^2}$$

where V_i^p is the p-th quantitative descriptors of the amino acid in i-th sequence. N is the number of sequences in the multiple sequence alignment. Higher average distance indicates more diversity in terms of physical-chemical properties at that column. The physical-chemical distance is converted to a physical-chemical similarity S by

$$S = \frac{N_{no\ gap}}{N} \exp(-0.1D)$$

where S is scaled between 0 and 1 with 1 indicating absolute conservation (identical column) and smaller numbers indicating more diversity. The similarity is also negatively affected when there are gaps in the alignment.

InterProSurf

InterProSurf is web server develop by our group to predict protein-protein interfaces (Negi and Braun, 2007; Negi et al., 2007). Previous studies showed that it is able to accurately predict the residues involved in the interface of the dimeric structure of ATPase and dimer interface of human hepatitis B virus capsid protein (Negi and Braun, 2007). Recently we have designed a variant of this software specifically for predicting antibody binding epitopes. The prediction of epitopes is based on surface exposed area and propensity of amino acids being in antibody binding sites. The epitope propensity scores (Figure 4.4) were derived from an analysis of 90 known X-ray crystal structures of antibody-protein complexes (unpublished data provided by Dr. S.S.Negi). The software scans through the surface of a 3D structure and searches for patches including one center residue and neighboring surface residues within certain radius. Then score of one patch is calculated by averaging the epitope propensity scores of all amino acids in the patch with weighting factors proportional to the solvent accessible surface area. The patches are ranked by their scores and the top patches are reported as potential antibody binding epitopes. The solvent accessible surface areas were calculated by the *GetArea* (Fraczkiewicz, 1998). *InterProSurf* is available at

http://curie.utmb.edu/patchAntigen.html

In this study, surface exposed residues in the viral envelope proteins were analyzed with the 3D structures of VEEV and EEEV and potential conformational epitopes were predicted by. The top 5% of the predicted epitopes for EEEV were used for further design of hybrid proteins.

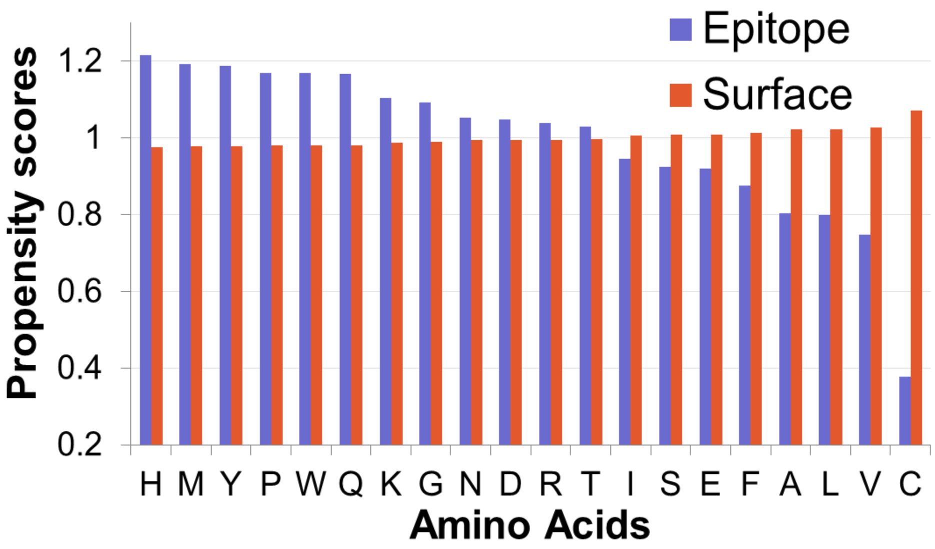


Figure 4.4 Propensity scores of epitope and surface for 20 amino acids.

(Negi & Braun, unpublished results)

Construction of Hybrid Viruses

Viruses containing the hybrid E2 proteins were constructed in Dr. Naomi Forrester's group at UTMB. Two regular primers were designed to anneal to the start and end of E2 gene. Additional primers were designed to include the point mutations for each of the hybrids (1-4). Site directed mutagenesis was performed with one regular primer and one mutation primer to obtain a segment containing the mutations. The segments were then joined by ligation PCR. Multiple rounds of ligation PCR may be required depending on the number of mutations. Finally, the E2 gene containing mutations of different hybrid proteins were inserted into the Tc-83 vector between restriction sites PspOMI and SgrAI resulting in complete genome of TC-83 with the mutations.

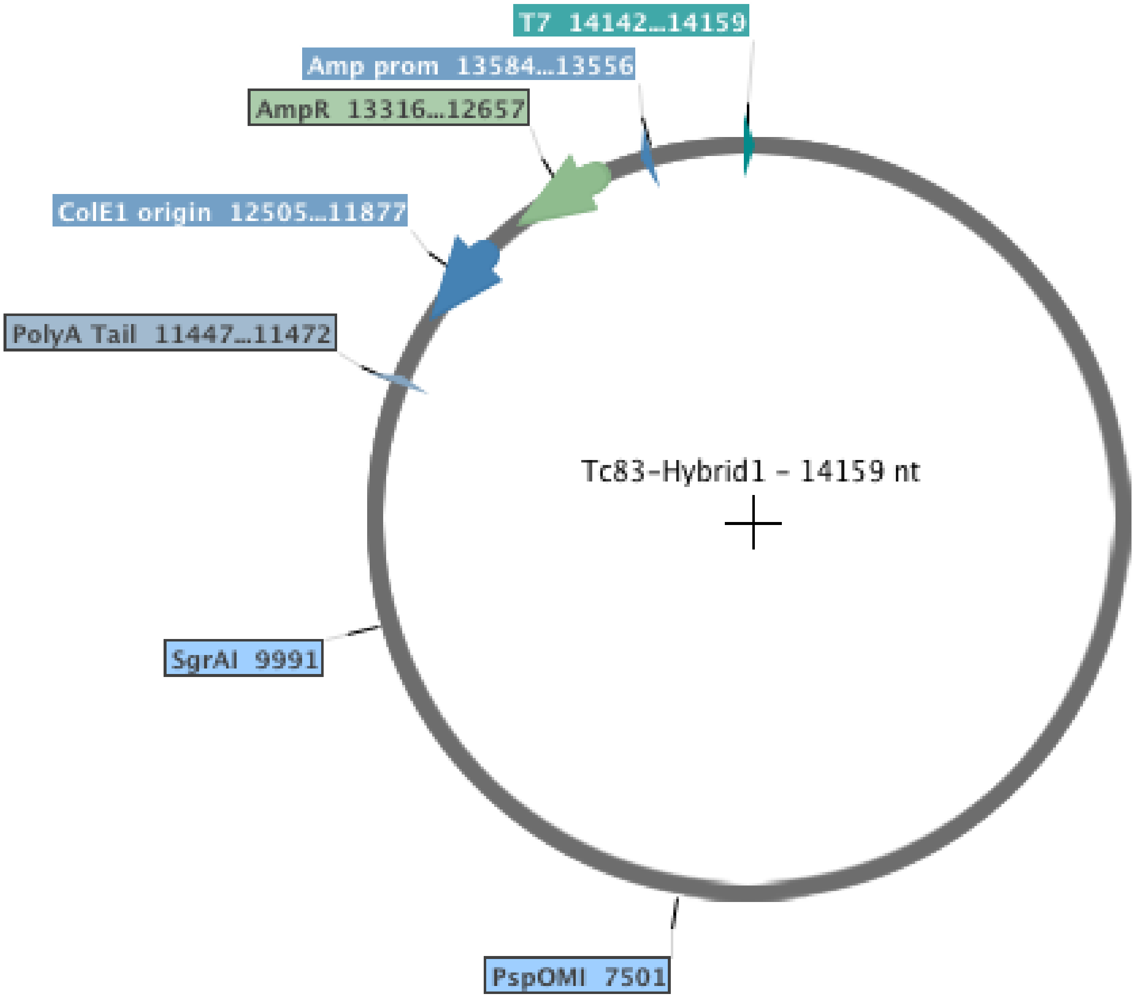


Figure 4.5 Plasmid map for expressing hybrid proteins

4.3 Results

Sequences of VEEV and EEEV E2 protein

The genomic sequences of VEEV and EEEV were obtained from Dr. Naomi Forrester's group as a result of high-throughput sequencing of wild-caught animal hosts and mosquito vectors. The sequences were translated to amino acid sequences for further analysis. Since domains A and B of envelope protein E2 are the most exposed domains on the surface, and they play important roles in receptor-binding, they are considered to be the main targets of the neutralizing antibodies (Metz and Pijlman, 2011; Vaney et al., 2013). Thus, the sequences and structures of the ectodomains of E2 protein (aa 1-340) are used to design multivalent vaccines.

A total number of 108 VEEV E2 sequences and 45 EEEV E2 sequences were collected and multiple sequence alignments were generated for each strain with *ClustalW2* (Larkin et al., 2007). The sequences within each species are highly similar with average sequence identity of 92.13% for VEEV and 92.74% for EEEV, but the sequence identity between two species is only ~60%. The conservation of E2 protein within each species suggests it is feasible to identify conserved epitopes for antibodies, while the diversity between species brings the challenge of designing multivalent vaccines. The sequence identity matrix of VEEV E2 indicates clearly separated subtypes of IE and ID (Figure 4.6), which are more common in wild hosts in enzootic cycle (Weaver et al., 2012). Meanwhile, EEEV E2 proteins also show clusters of subtypes I-IV (Figure 4.7). To avoid the bias caused by redundant sequences, I clustered the sequences at 99% identity by CD-HIT (Li and Godzik, 2006) and selected representative sequences (red label) for the identification of conserved and variable regions. Average PCP distances were calculated for each columns of the alignment where higher value indicates more diversity at that column. (Figure 4.9 a)

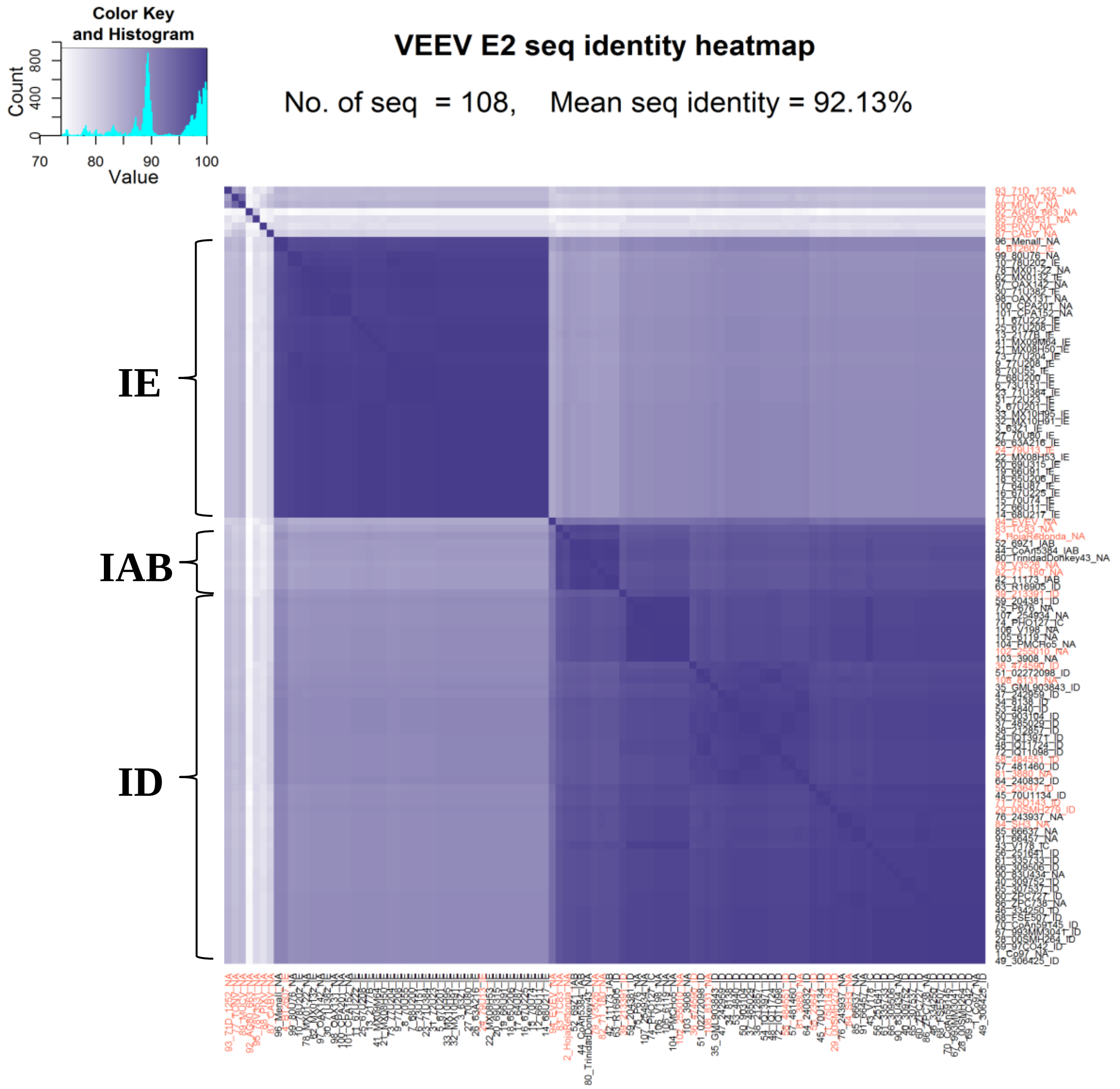


Figure 4.6 Sequence identity heat map of the E2 protein of VEEV.

108 VEEV E2 sequences are included in the heat map with average sequence identity of 92.13%. High sequence identity is indicated by dark blue while light blue shows low sequence identity. The histogram of sequence identities is shown in cyan in the top left corner. The clusters of subtypes are labeled on the left. Redundant sequences with >=99% sequence identity are removed by CD-HIT, the remaining reprehensive sequences are labeled in red.

:

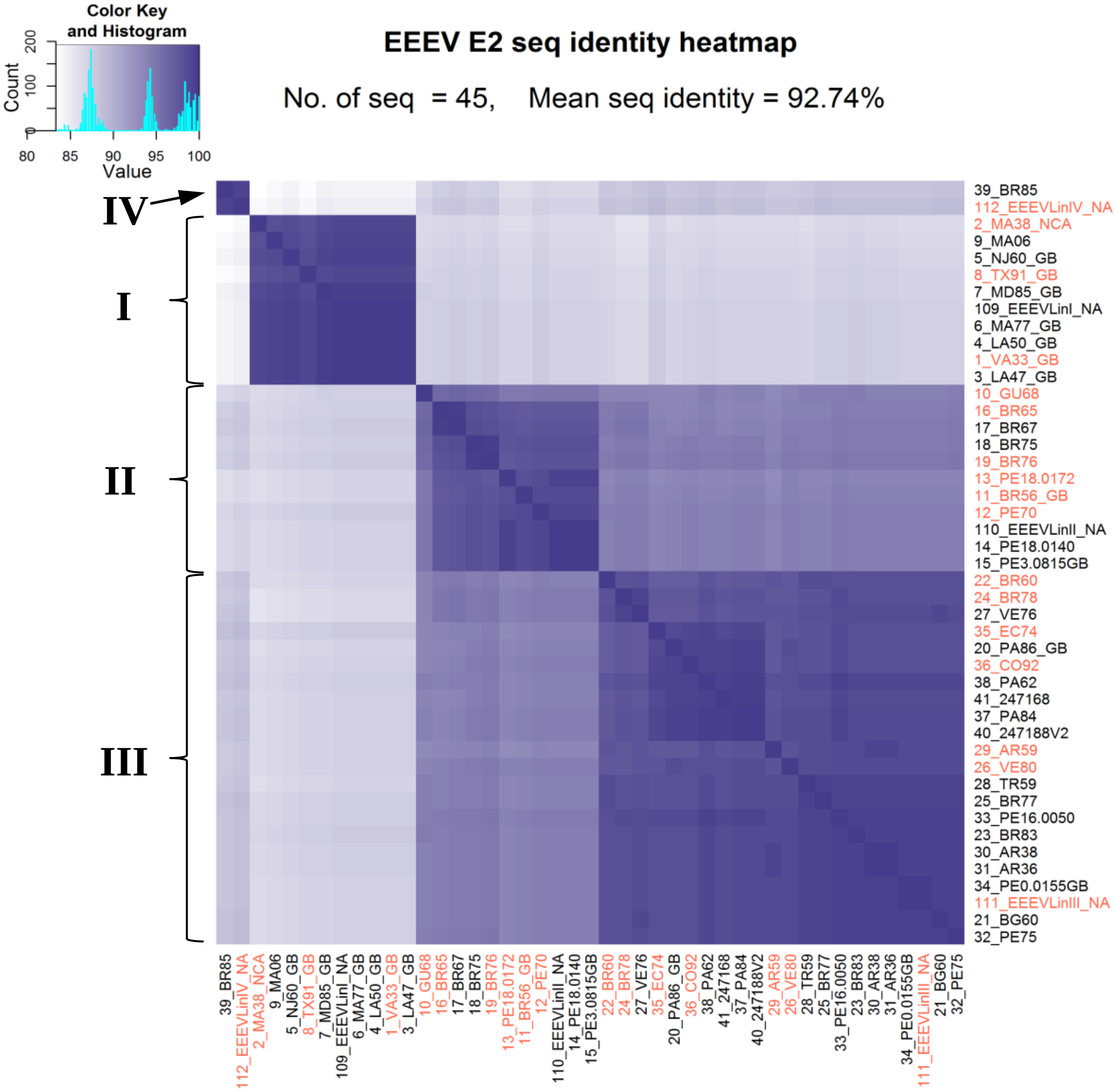


Figure 4.7 Sequence identity heat map of the E2 protein of EEEV

45 EEEV E2 sequences are included in the heat map with average sequence identity of 92.74%.

Predicted epitopes on E2 protein of VEEV

In addition to the sequence analysis, I also used the structures of VEEV and EEEV for predicting potential antibody binding epitopes. The prediction was done with InterProSurf at patch radius of 8.0 Å for the high resolution cryo-EM structure of VEEV strain TC-83 (PDB: 3J0C) (Zhang et al., 2011). The predicted epitopes were mapped on the structure with red as high epitope scores and blue as low scores. A comparison with the previously reported experimental epitopes (aa 115-119, 182-207) (Hunt et al., 2010; Roehrig and Mathews, 1985) indicates that the predicted epitopes partially overlap the experimental epitopes (Figure 4.8 A, C).

For EEEV, since there are no experimental structures available, a 3D homology model was built for the ectodomains of strain VA33 (aa 1-338) using our modeling program MPACK and the crystal structure of Chikungunya virus (PDB: 3N40) (Voss et al., 2010) as template. The backbone root-mean-square-deviation (RMSD) of 0.7 Å indicates the model is accurate and reliable for the prediction of epitopes. As expected, several epitopes with high scores are located on the most exposed domain B. No experimental epitopes of EEEV are available for comparison, but the predicted epitopes match with some of the experimental results for Chikungunya virus, such as the human IgG binding sites E2EP3 (Kam et al., 2014; Kam et al., 2012) and the conformational epitope of Fab fragments of the neutralizing mouse MAbs, CHK9, m242 (Sun et al., 2013) (Figure 4.8 B, D).

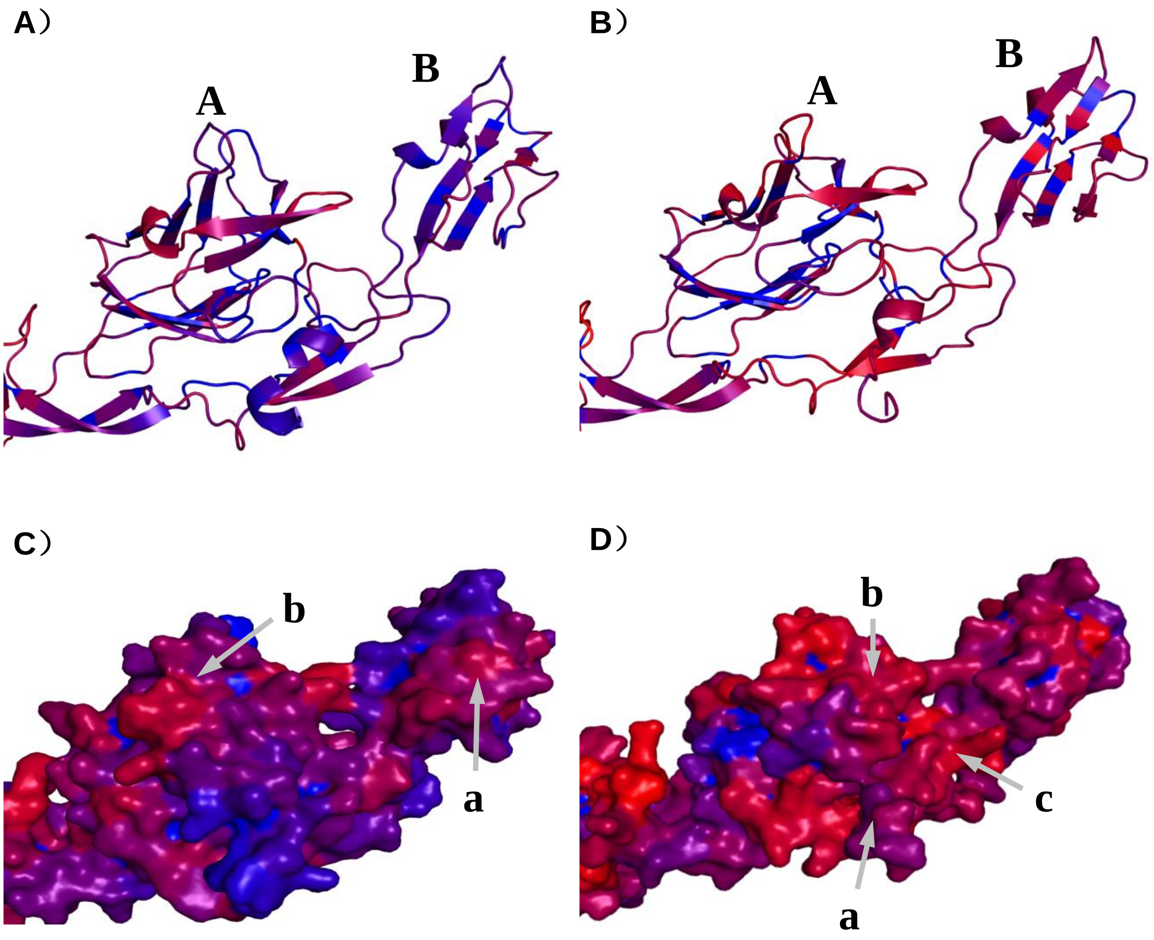


Figure 4.8 Predicted antibody binding sites for E2 protein of VEEV and EEEV

A), C), predicted epitopes were mapped on VEEV TC83. High score epitopes are in red while low scores in blue. Domains A and B are labeled in A). Overlaps with experimental epitopes are marked by a (aa 182-207) and b (aa 115-119) (Hunt et al., 2010; Roehrig and Mathews, 1985) in C).

B), D), predicted epitopes were mapped on EEEV VA33. a, b and c match with experimental epitopes E2EP3 (Kam et al., 2014; Kam et al., 2012), CHK9 and m242 (Sun et al., 2013) respectively.

Design of hybrid E2 protein

For the design of multivalent vaccines, we need to identify conserved epitopes, especially for EEEV where no experimental epitopes are available. I combined the sequences analysis of E2 protein and the prediction of epitopes to define the conserved epitopes for EEEV. Firstly, the scores of the predicted epitopes were mapped on the reference sequence of EEEV VA33 with color codes indicating the scores (Figure 4.9 b, red: high score; blue: low scores; blank: buried residues), and the residues in the top 5% scoring epitopes were marked by reference residues (Figure 4.9 b, red residues). Then, the PCP similarity of each column in the epitopes was calculated and the top epitopes were evaluated for their conservation by averaging the PCP similarities of all involving columns. A cutoff value of 0.9 was set for PCP similarities to identify the conserved, high-scoring epitopes (Figure 4.9 c, black residues).

Our strategy of developing multivalent vaccines against both VEEV and EEEV is to design E2 proteins that represent the immune features of both strains (Figure 4.10). Since the VEEV vaccine strain TC-83 is well studied with available 3D structures and experimental epitopes, I used TC-83 as backbone and introduced conserved epitopes from EEEV for the design of hybrid E2 proteins. Additionally, to ensure the hybrid proteins form stable folds without significant conformational changes, only the variable regions of backbone TC-83 could be mutated. Thus, I identified a list of residues that are conserved and with high epitope scores in EEEV, while their corresponding positions in VEEV are composed of rather diverse residues with similarity scores below average (Figure 4.9 d, yellow residues). The two predicted epitopes for EEEV (patch centers: 162, 233) and experimental epitopes for VEEV were mapped on the structure of VEEV TC-83

(Figure 4.11 A) and on a spike complex (Figure 4.11 B, C). Their locations are fully exposed on the surface of virus where they can be easily accessed by antibodies.

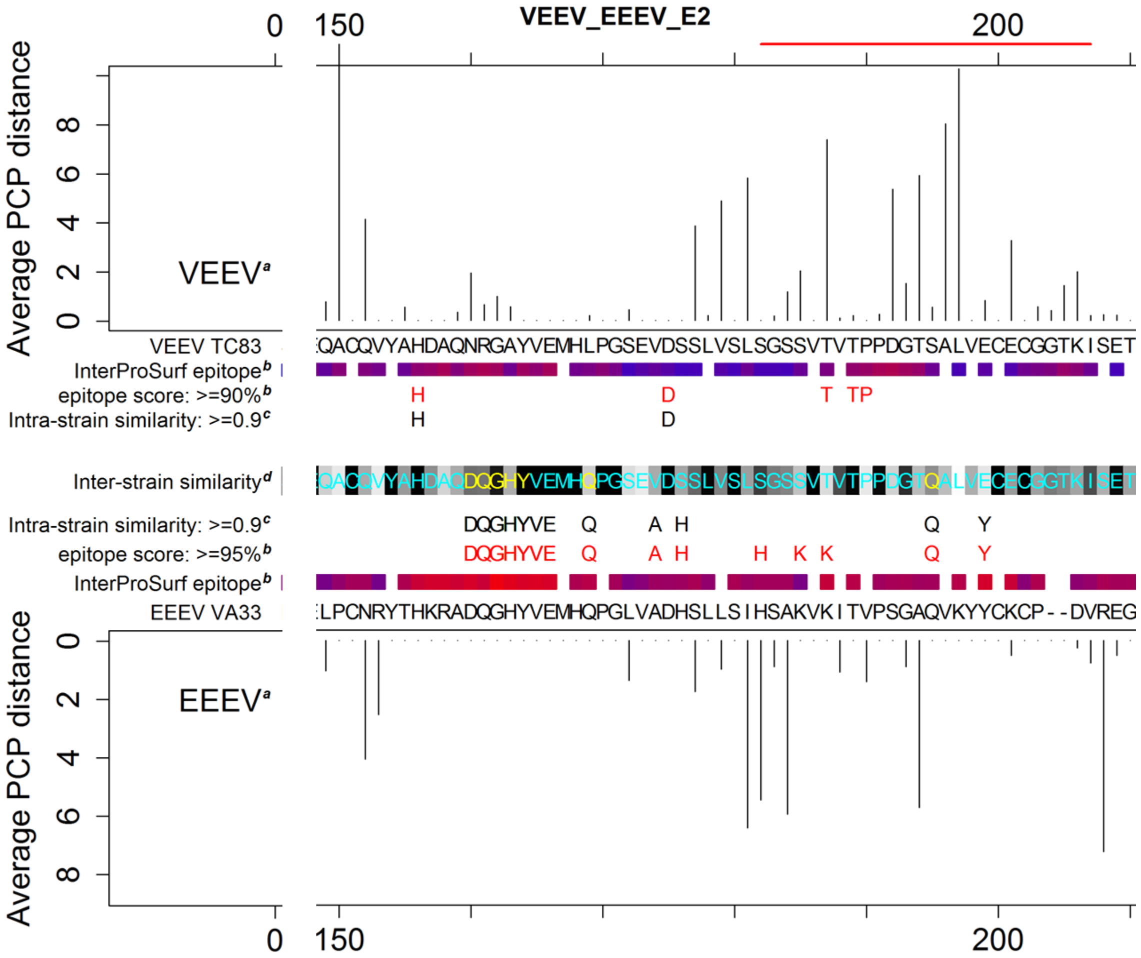


Figure 4.9 Mapping conserved residues and predicted epitopes on VEEV and EEEV. The upper panel is the information for VEEV and lower for EEEV. Only aa 150-250 is shown. a). The average PCP distance (vertical lines) for each column was calculated within each strain. Higher distance indicates more diverse residues.

b). The predicted epitopes were mapped on reference sequence with red for high score, blue for low scores and blank for buried residues. Epitopes with top 10%(VEEV) or 5% (EEEV) high scores were listed as red residues.

c). High-score epitopes with PCP similarities >= 0.9 are listed as black residues.

d). The inter-strain similarity was shown as grey scale bar with black for high similarity.

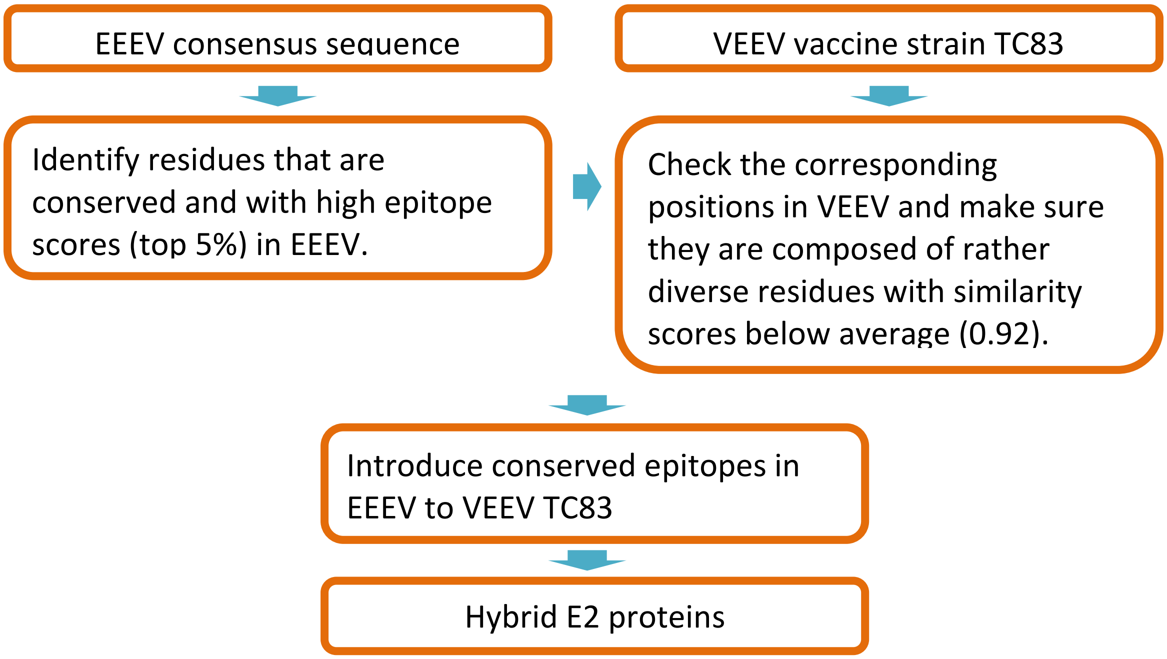
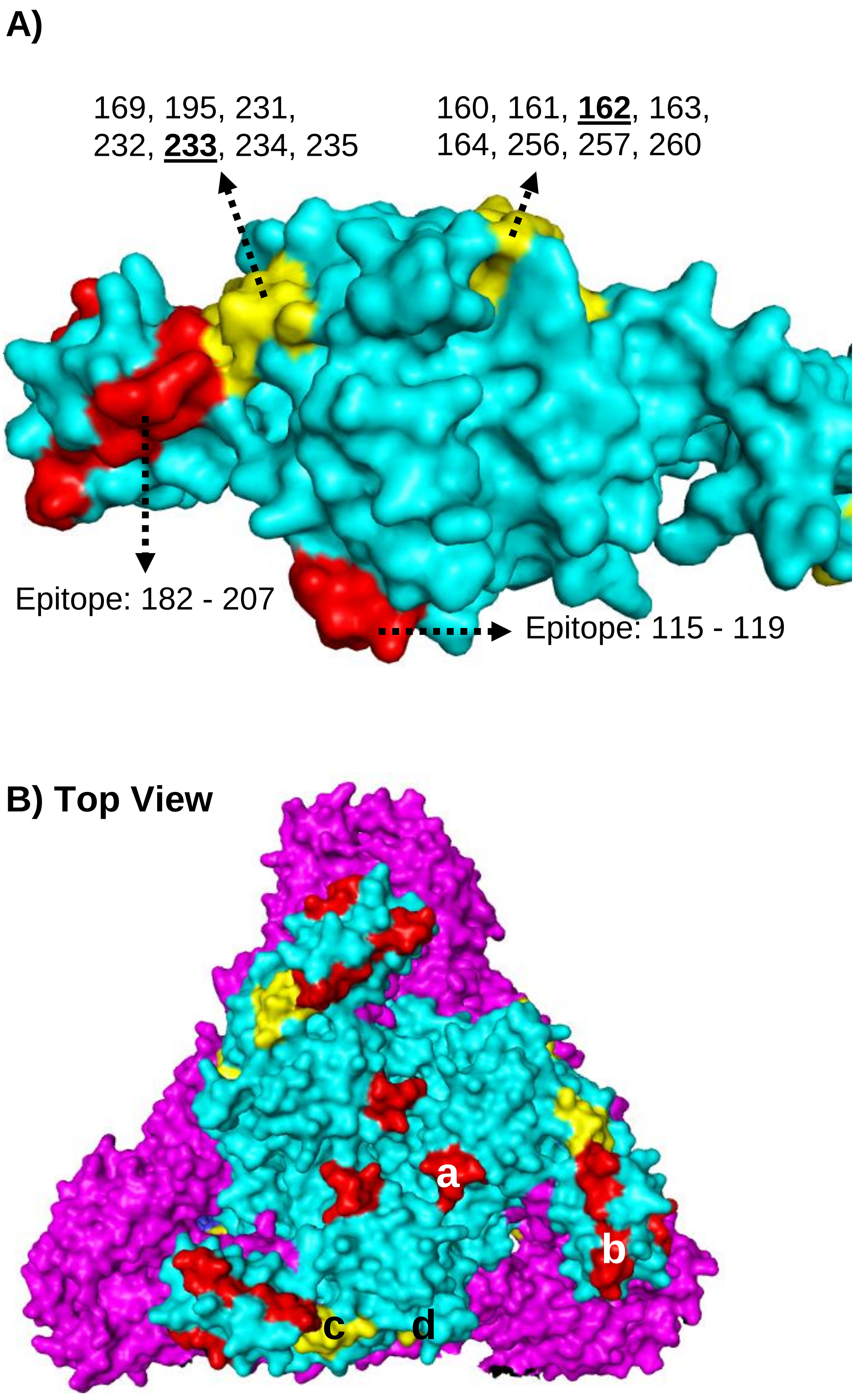

EEEV consensus sequence
VEEV vaccine strain TC83
Identify residues that are conserved and with high epitope scores (top 5%) in EEEV.
Check the corresponding positions in VEEV and make sure they are composed of rather diverse residues with similarity scores below average (0.92).
Introduce conserved epitopes in EEEV to VEEV TC83
Hybrid E2 proteins

Figure 4.10 The strategy to design hybrid E2 proteins



A)

B) Top View

169, 195, 231,
232, **233**, 234, 235
160, 161, **162**, 163,
164, 256, 257, 260
Epitope: 182 - 207
Epitope: 115 - 119
a
b
c
d

C) Side View

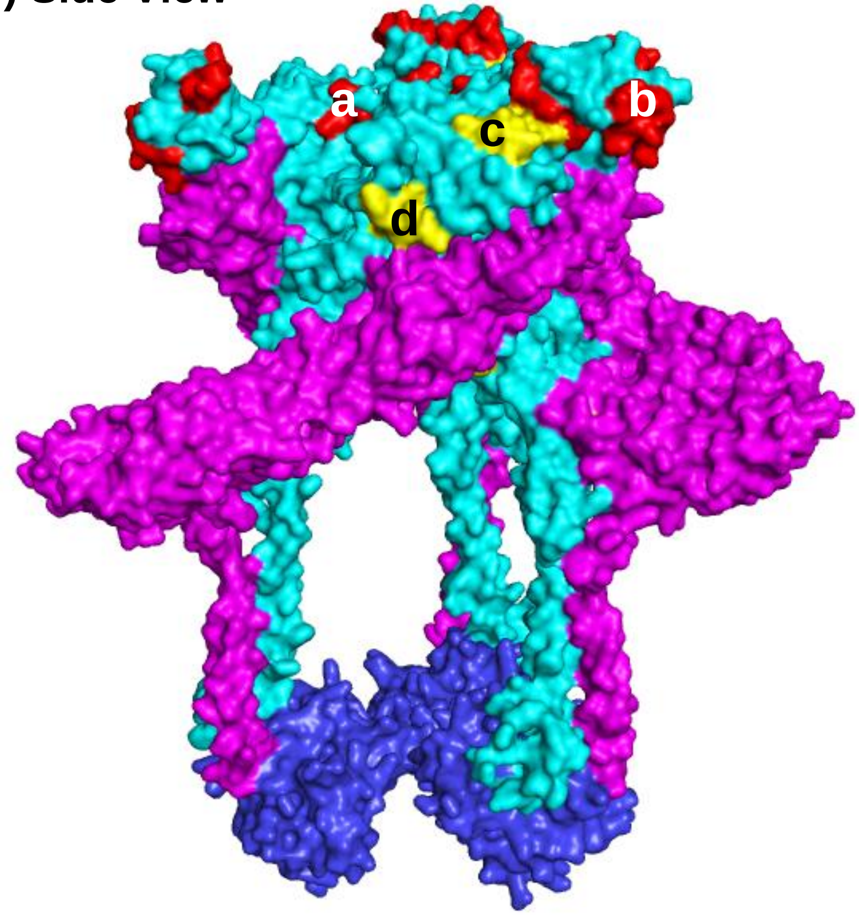


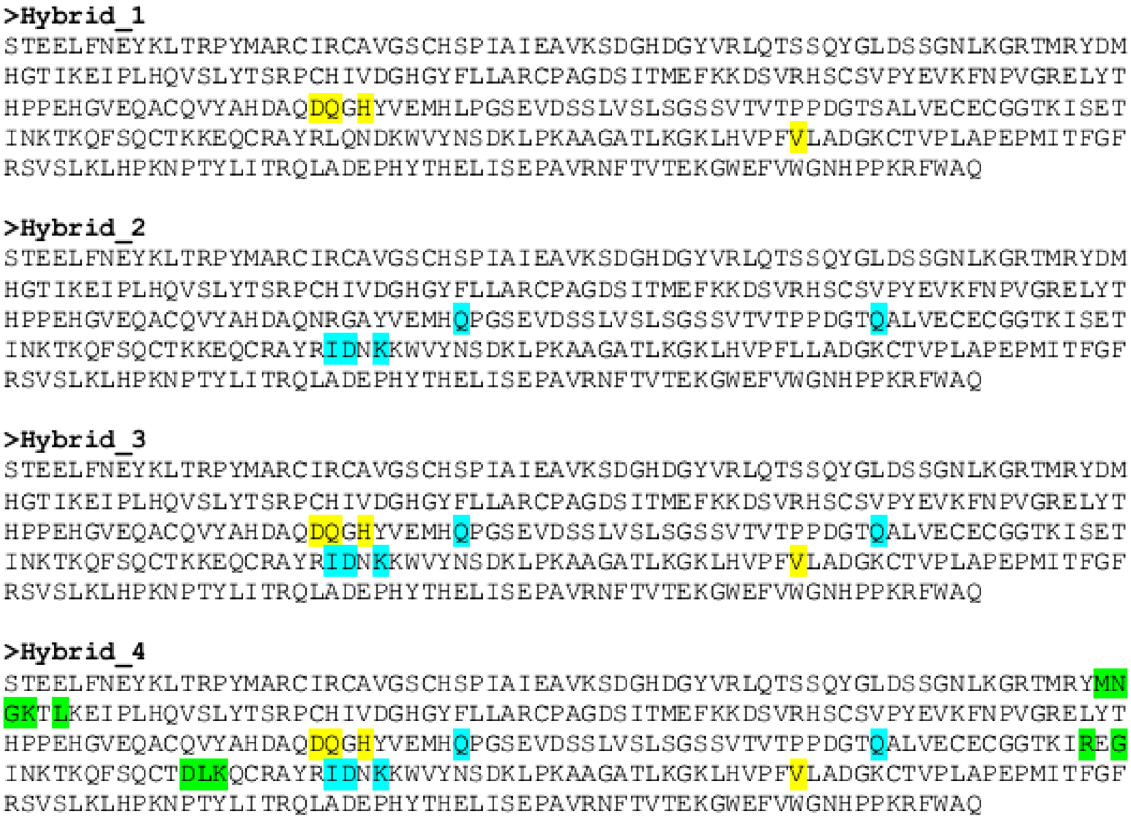
Figure 4.11 Experimental epitopes for VEEV and predicted epitopes for EEEV.

A). The epitopes on the domains A (right) and B (left) of VEEV TC83. Red: experimental epitope (aa 115-119, 182-207) (Hunt et al., 2010; Roehrig and Mathews, 1985). Yellow: top 5% predicted epitopes from EEEV. The centers of patches are in bold. B), C). Top and side views of the epitopes in the complex of a spike. Magenta: E1; Cyan: E2; Blue: capsid. a, VEEV epitope aa 115-119; b, VEEV epitope aa 182-207. c, EEEV epitope with center of aa233; d, EEEV epitope with center of aa 162.

Homology models of Hybrid E2 proteins

Based on the previous strategy, I designed four hybrid E2 proteins to represent the immune features of both VEEV and EEEV (Figure 4.12). Hybrid 1 and 2 were constructed by inserting each of the top 5% predicted epitopes from EEEV, while hybrid 3 contains both epitopes. To make sure no potential epitopes are eliminated because of the high cutoff values, I slightly released the cutoff of epitope scores for EEEV and included another two conserved epitopes that are among the top 8% scores in hybrid 4. The complete sequences of the four hybrid E2 proteins are shown in the following figure (Figure 4.12) and the mutation sites are listed in the table (Table 4.1).

To evaluate the stability of these hybrid E2 proteins, I built homology models for all four constructions using MPACK and the high resolution cryo-EM structure of VEEV TC-83 (PDB: 3J0C) (Zhang et al., 2011) as template. The comparison between the hybrid E2 proteins with original TC-83 shows they overlap with each other in most regions with small backbone RMSD value (Figure 4.13). The energy minimization of the models using FANTOM (Schaumann et al., 1990) shows low energies for all hybrid proteins, suggesting that they are able to form stable 3D structures (Table 4.2). I further assessed the quality of the models using three other softwares, QMEAN (Benkert et al., 2011), ProSA (Wiederstein and Sippl, 2007), and Verify_3D (Bowie et al., 1991). The results indicate that the scores of homology models are similar to the experimental structure of VEEV TC-83 (Table 4.2).



```
>Hybrid_1
STEELFNEYKLTRPYMARCIRCAVGSCHSPIAIEAVKSDGHDGYVRLQTSSQYGLDSSGNLKGRTMRYDM
HGTIKEIPLHQVSLYTSRPCHIVDGHGYFLLARCPAGDSITMEFKKDSVRHSCSVPYEVKFNPVGRELYT
HPPEHGVEQACQVYAHDAQDQGHYVEMHLPGSEVDSSLVSLSGSSVTVTPPDGTSALVECECGGTKISET
INKTKQFSQCTKKEQCRAYRLQNDKWVYNSDKLPKAAGATLKGKLHVPFVLADGKCTVPLAPEPMITFGF
RSVSLKLHPKNPTYLITRQLADEPHYTHELISEPAVRNFTVTEKGWEFVWGNHPPKRFWAQ

>Hybrid_2
STEELFNEYKLTRPYMARCIRCAVGSCHSPIAIEAVKSDGHDGYVRLQTSSQYGLDSSGNLKGRTMRYDM
HGTIKEIPLHQVSLYTSRPCHIVDGHGYFLLARCPAGDSITMEFKKDSVRHSCSVPYEVKFNPVGRELYT
HPPEHGVEQACQVYAHDAQNRGAYVEMHQPGSEVDSSLVSLSGSSVTVTPPDGTQALVECECGGTKISET
INKTKQFSQCTKKEQCRAYRIDNKKWVYNSDKLPKAAGATLKGKLHVPFLLADGKCTVPLAPEPMITFGF
RSVSLKLHPKNPTYLITRQLADEPHYTHELISEPAVRNFTVTEKGWEFVWGNHPPKRFWAQ

>Hybrid_3
STEELFNEYKLTRPYMARCIRCAVGSCHSPIAIEAVKSDGHDGYVRLQTSSQYGLDSSGNLKGRTMRYDM
HGTIKEIPLHQVSLYTSRPCHIVDGHGYFLLARCPAGDSITMEFKKDSVRHSCSVPYEVKFNPVGRELYT
HPPEHGVEQACQVYAHDAQDQGHYVEMHQPGSEVDSSLVSLSGSSVTVTPPDGTQALVECECGGTKISET
INKTKQFSQCTKKEQCRAYRIDNKKWVYNSDKLPKAAGATLKGKLHVPFVLADGKCTVPLAPEPMITFGF
RSVSLKLHPKNPTYLITRQLADEPHYTHELISEPAVRNFTVTEKGWEFVWGNHPPKRFWAQ

>Hybrid_4
STEELFNEYKLTRPYMARCIRCAVGSCHSPIAIEAVKSDGHDGYVRLQTSSQYGLDSSGNLKGRTMRYMN
GKTLKEIPLHQVSLYTSRPCHIVDGHGYFLLARCPAGDSITMEFKKDSVRHSCSVPYEVKFNPVGRELYT
HPPEHGVEQACQVYAHDAQDQGHYVEMHQPGSEVDSSLVSLSGSSVTVTPPDGTQALVECECGGTKIREG
INKTKQFSQCTDLKQCRAYRIDNKKWVYNSDKLPKAAGATLKGKLHVPFVLADGKCTVPLAPEPMITFGF
RSVSLKLHPKNPTYLITRQLADEPHYTHELISEPAVRNFTVTEKGWEFVWGNHPPKRFWAQ
```

Figure 4.12 Sequences of designed hybrid E2 proteins.

Yellow, Cyan: conserved epitopes of top 5% high score for EEEV

Green: additional conserved epitopes of top 8% high score for EEEV

Table 4.1 Mutation sites in hybrid E2 proteins

Hybrid Proteins	Hybrid_1	Hybrid_2	Hybrid_3	Hybrid_4	
No. of mutations	4	5	9	19	
Sites of mutations	N160D		N160D	N160D	D69M
	R161Q		R161Q	R161Q	M70N
	A163H		A163H	A163H	H71G
	L260V		L260V	L260V	G72K
		L169Q	L169Q	L169Q	I74L
		S195Q	S195Q	S195Q	S208R
		L231I	L231I	L231I	T210G
		Q232D	Q232D	Q232D	K222D
		D234K	D234K	D234K	K223L
					E224K

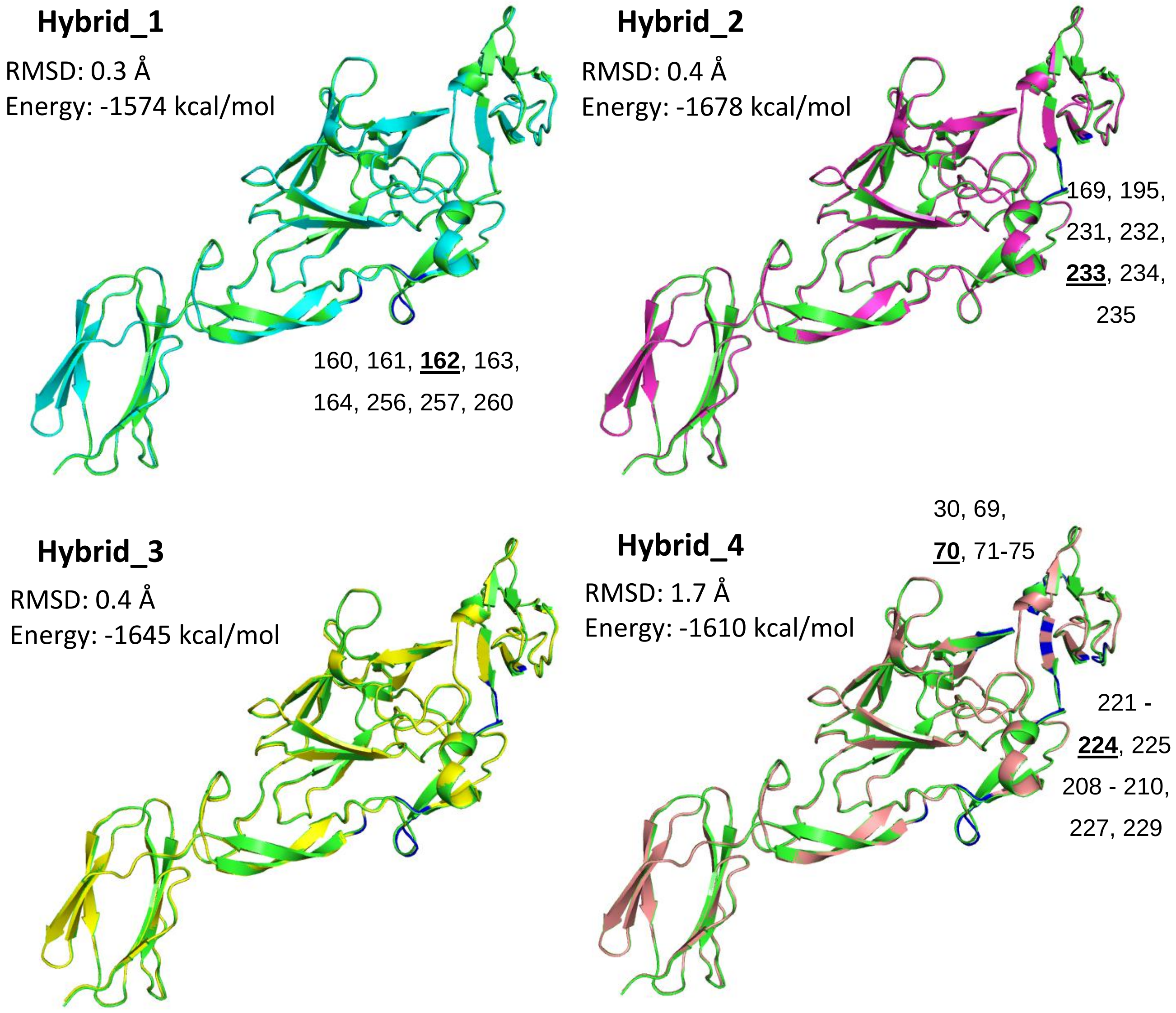


Figure 4.13 Homology models of the hybrid E2 proteins.

Green: cryo-EM structure of VEEV TC-83; Cyan: hybrid 1; Magenta: hybrid 2; Yellow: hybrid 3; Pink: hybrid 4. The mutated residues are marked in blue. The predicted epitopes of EEEV are labeled by their residues indexes. The RMSD and minimized energies are shown on the top left corner of each structure.

Table 4.2 Assessment of the quality of the models

Models	Minimized energy (kcal/mol)	RMSD (Å)	QMEAN Score (Z- score)	ProSA Z-score	Verify_3D Average Score > 0.2 (%)
VEEV_TC83	-	-	0.632 (-1.67)	-6.72	86.4 (pass)
Hybrid_1	-1574	0.3	0.621 (-1.80)	-6.70	86.2 (pass)
Hybrid_2	-1678	0.4	0.638 (-1.60)	-6.61	85.6 (pass)
Hybrid_3	-1645	0.4	0.637 (-1.62)	-6.55	85.0 (pass)
Hybrid_4	-1610	1.7	0.631 (-1.69)	-6.44	84.8 (pass)

4.4 Discussion

Encephalitic alphaviruses are important human and veterinary pathogens that cause massive outbreaks in Central, South America and the U. S. The epidemics result in great medical burdens and economic devastations. VEEV and EEEV are among the most virulent members of the genus that may cause severe symptoms of neurological sequelae or death. The viruses have been developed as biological weapons due to their debilitating symptoms and high efficiency of aerosol infections (Bronze et al., 2002; Hawley and Eitzen, 2001). Therefore, there is a pressing need for vaccines against a broad spectrum of alphaviruses to limit massive outbreaks and defend attacks of biological weapons.

Great efforts have been spent to develop vaccines against encephalitic alphaviruses. The conventional methods often involve isolation of naturally attenuated viruses or viruses that are attenuated by random evolving mutations when passaging in cell culture. However, the vaccines from these methods are at high risks of residual pathogenicity and may revert to virulence (Metz and Pijlman, 2011; Paessler and Weaver, 2009). The formalin-inactivated VEEV vaccines used in middle 20 centuries were shown to be effective in horses while causing very mild symptoms (Randall et al., 1949). But the isolation of residual live viruses in the vaccine indicates its potential virulence (Sutton and Brooke, 1954). The attenuated VEEV vaccine TC-83 was widely used in equids during outbreaks after 1960s and experiments showed its ability to induce neutralizing antibodies in human (Walton et al., 1972); however, there are only two mutations in TC-83 comparing to its virulent ancestor, one in 5'-noncoding region and the other one (T120R) in E2 envelope glycoprotein, suggesting the attenuation may be incomplete (Kinney et al., 1993). Additional studies showed that the vaccine caused significant side effects in many volunteers (Paessler and Weaver, 2009; Pittman et al., 1996) and it is also capable of infecting mosquito vectors and spreading to other hosts (Pedersen et al., 1972). Thus, it is limited to the vaccination of lab personnel at risk of occupational exposure. Further attenuation of TC-83 was achieved by inserting RNA elements that are only functional in vertebrate cells to prevent the replication of viruses in mosquito vectors (Volkova et al., 2008).

In addition to traditional methods, genetic engineering has been applied to the rational design of vaccines, where the vaccines contain defined attenuating mutations or large segments of genes from other nonpathogenic viruses. Pratt et.al. designed a vaccine strain, V3526, by deleting the cleavage signal (residues 56–59) of PE2, the precursor to mature E2 glycoprotein, and one point mutation at envelope protein E1 253 (Hart et al., 2000; Pratt et al., 2003). The vaccine was tested in multiple animal models and shown to be viable, immunogenic against subtype IAB and less neurovirulent than TC-83 (Fine et al., 2008; Ludwig et al., 2001), making it a promising candidate. Another approach is to design recombinants of virulent and non-virulent alphaviruses. Chimeric SIN/VEE viruses were constructed using the non-structural protein genes of the relatively non-virulent Sindbis virus (SINV) to control the replication and transcription of the structural proteins of the virulent VEEV (Paessler et al., 2003). The studies in mice demonstrated that the chimeric vaccines are highly efficacious against the challenge of pathogenic VEEV without causing any detectable clinical disease (Ni et al., 2007). Another attempt by Dupuy et al. employed libraries of recombined envelope proteins from VEEV, EEEV and WEEV. By high-throughput screening, they identified recombinants that have improved immunogenicity and protective efficacy against VEEV subtype IAB (Dupuy et al., 2009). Despite the significant progress, these designs of vaccines are still in the process of evaluating their viability, safety and immunogenicity; further clinical tests in human are required. Additionally, these vaccines only induce neutralizing antibodies against one or two subtypes of VEEV, vaccines that protect against multiple strains of encephalitic alphaviruses is still a challenge due to the diversity of their structural protein sequences. In case of a sudden outbreak with an emerging pathogen, there is not enough time to develop a vaccine against the circulating strain, thus a broad protecting vaccine is needed to prepare for such an event. Currently, no vaccines are licensed in the U.S. for mass immunization during outbreaks (Metz and Pijlman, 2011; Paessler and Weaver, 2009).

We presented an alternative strategy for rational design of multivalent vaccines against VEEV and EEEV. Instead of rather randomly screening for mutation sites that are involved in antibody binding, our method used computational methods for the prediction of antibody binding epitopes in the 3D structures of viruses’ envelope proteins. These conformational epitopes are composed of critical residues that may be distant in sequence but close in space, which are difficult to detect by experimental method of screening escape mutations. The combination of sequential and conformational information indicates the conserved regions that are important for functions or antibody binding and variable regions that are less vulnerable to genetic modification. The resulting hybrid E2 proteins are promising candidates for vaccines that induce immune responses against both VEEV and EEEV.

4.5 Conclusions and future studies

In this project, I analyzed the conserved and variable regions in the sequences of the envelope protein E2 of encephalitic alphaviruses VEEV and EEEV. Potential antibody binding epitopes were predicted using their 3D structures. I combined the sequential and conformational analysis, and designed hybrid E2 proteins that represent the immune features of both VEEV and EEEV by introducing conserved epitopes in EEEV to the VEEV vaccine strain TC-83. The hybrid E2 proteins are expected to induce immune responses against both VEEV and EEEV.

Experimental validations of the hybrid proteins are in process in collaboration with Dr. Naomi Forrester. For future studies, we will finish the construction of the genomes of TC-83 containing these hybrid proteins. The viability of the viruses will be tested to ensure that the mutations do not result in mis-folded envelope proteins. Then we will vaccinate mice with them and test for neutralizing antibodies against VEEV and EEEV.

CHAPTER 5: SUMMARY and Future activities

The physical-chemical properties (PCP) of amino acids provide the essential information for the structures and functions of proteins. Critical residues that are either continuous in sequence or close in space are more likely to be conserved in their PCP. Therefore identification of these residues would significantly deepen our understanding of the structures and functions of proteins. In our studies, we have successfully demonstrated that PCP and conformational motifs identified by our method have significant impacts on various functions and properties of proteins.

We identified sequence motifs that are specific to Immunoglobulin (Ig) domains with different mechanical strength based on PCP. The motifs locate in the force-bearing regions that are critical to mechanical strength according to previous studies. Our experimental study with AFM demonstrated that the strength of mechanically weak Ig domains is enhanced by introducing specific motifs from stronger domains, indicating the importance of the motifs in maintaining the mechanical stability.

We identified allergen-specific motifs for major allergens in the top 17 Pfam-A families based on PCP and developed a novel scoring method to evaluate potential allergenicity of query sequences. The scoring method is able to distinguish non-allergenic sequences and allergenic sequences in these families at low false positive rates, providing a promising approach in addition to current methods for fast initial evaluation of allergenicity. Validation with the example sequences of major peanut allergens, Pectate lyase (PF00544) and allergens in PF00407 demonstrated that the allergen-specific motifs overlap with experimental IgE epitopes, distinguish homologous non-allergenic sequences and provide insight in the cross-reactivity of allergens.

We analyzed the sequences and structures of envelope proteins E2 of VEEV and EEEV, and identified conserved antibody binding epitopes in E2 protein of EEEV based on PCP and InterProSurf. We designed hybrid E2 proteins that represent the immune features of both VEEV and EEEV by introducing conserved epitopes in EEEV to the VEEV vaccine strain TC83. We are now in the process of constructing the virus genome with the hybrid proteins and validating the immune responses of mice against them. The virus-like particles containing these hybrid proteins are expected to be new candidates for developing safe, immunogenic and efficacious vaccines against multiple species of encephalitis alphaviruses.

The usefulness of PCP motifs in the understanding of the structure and function of proteins, and in the studies of protein engineering, drug development and vaccine design have been demonstrated by the previous projects in my research. As a general motif mining method, PCPMer has the potential to be applied to proteins with diverse functions and structures to identify critical residues with conserved physical–chemical properties. In addition to the current mining method for PCP motifs that mainly focuses on residues continuous in sequence, the structural information will be included for the further development of this method. For example, conformational PCP motifs could be generated by utilizing the 3D structures and identifying clusters of PCP-conserved residues that are within certain distance in space.

REFERENCES

AAAAI. 2010. Allergy Statistics.

Abagyan, R.A., and S. Batalov. 1997. Do aligned sequences share the same fold? *J Mol Biol*. 273:355-368.

Aceituno, E., V. Del Pozo, A. Minguez, I. Arrieta, I. Cortegano, B. Cardaba, S. Gallardo, M. Rojo, P. Palomino, and C. Lahoz. 2000. Molecular cloning of major allergen from Cupressus arizonica pollen: Cup a 1. *Clin Exp Allergy*. 30:1750-1758.

Al-Muhsen, S., A.E. Clarke, and R.S. Kagan. 2003. Peanut allergy: an overview. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne*. 168:1279-1285.

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol*. 215:403-410.

Attwood, T.K., A. Coletta, G. Muirhead, A. Pavlopoulou, P.B. Philippou, I. Popov, C. Roma-Mateo, A. Theodosiou, and A.L. Mitchell. 2012. The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. *Database : the journal of biological databases and curation*. 2012:bas019.

Bailey, T.L., M. Boden, F.A. Buske, M. Frith, C.E. Grant, L. Clementi, J. Ren, W.W. Li, and W.S. Noble. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 37:W202-208.

Bailey, T.L., and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*. 2:28-36.

Bairoch, A., B. Boeckmann, S. Ferro, and E. Gasteiger. 2004. Swiss-Prot: juggling between evolution and stability. *Briefings in bioinformatics*. 5:39-55.

Beck, C.E., and R.W. Wyckoff. 1938. Venezuelan Equine Encephalomyelitis. *Science*. 88:530.

Benkert, P., M. Biasini, and T. Schwede. 2011. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. 27:343-350.

Berge TO, B.I., Tigertt WD. 1961. Attenuation of Venezuelan equine encephalomyelitis virus by in vitro cultivation in guinea pig heart cells. *Am. J. Hyg.* 73:209-218.

Best, R.B., S.B. Fowler, J.L. Herrera, A. Steward, E. Paci, and J. Clarke. 2003. Mechanical unfolding of a titin Ig domain: structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. *J Mol Biol*. 330:867-877.

Bjorklund, A.K., D. Soeria-Atmadja, A. Zorzet, U. Hammerling, and M.G. Gustafsson. 2005. Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics*. 21:39-50.

Bloomfield, S.F., R. Stanwell-Smith, R.W. Crevel, and J. Pickup. 2006. Too clean, or not too clean: the hygiene hypothesis and home hygiene. *Clin Exp Allergy*. 36:402-425.

Bolhaar, S.T., L. Zuidmeer, Y. Ma, F. Ferreira, C.A. Bruijnzeel-Koomen, K. Hoffmann-Sommergruber, R. van Ree, and A.C. Knulst. 2005. A mutant of the major apple allergen, Mal d 1, demonstrating hypo-allergenicity in the target organ by double-blind placebo-controlled food challenge. *Clin Exp Allergy*. 35:1638-1644.

Bonds, R.S., T. Midoro-Horiuti, and R. Goldblum. 2008. A structural basis for food allergy: the role of cross-reactivity. *Current opinion in allergy and clinical immunology*. 8:82-86.

Borgia, A., A. Steward, and J. Clarke. 2008. An effective strategy for the design of proteins with enhanced mechanical stability. *Angew Chem Int Ed Engl*. 47:6900-6903.

Bork, P., and E.V. Koonin. 1996. Protein sequence motifs. *Curr Opin Struct Biol*. 6:366-376.

Bowie, J.U., R. Luthy, and D. Eisenberg. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*. 253:164-170.

Bronze, M.S., M.M. Huycke, L.J. Machado, G.W. Voskuhl, and R.A. Greenfield. 2002. Viral agents as biological weapons and agents of bioterrorism. *Am J Med Sci*. 323:316-325.

Brutlag, A.B.-H.D. 2006. Sequence Motifs: Highly Predictive Features of Protein Function. *In* Feature Extraction: Foundations and Applications. I.G.M.N.S.G.L.A. Zadeh, editor. Springer Berlin Heidelberg. pp 625-645.

Bujnicki, J.M., A. Elofsson, D. Fischer, and L. Rychlewski. 2001. Structure prediction meta server. *Bioinformatics*. 17:750-751.

Bustamante, C., Y.R. Chemla, N.R. Forde, and D. Izhaky. 2004. Mechanical processes in biochemistry. *Annu Rev Biochem*. 73:705-748.

Cao, Y., T. Yoo, and H. Li. 2008. Single molecule force spectroscopy reveals engineered metal chelation is a general approach to enhance mechanical stability of proteins. *Proc Natl Acad Sci U S A*. 105:11152-11157.

Crampton, N., and D.J. Brockwell. 2010. Unravelling the design principles for single protein mechanical strength. *Curr Opin Struct Biol*. 20:508-517.

Cui, J., L.Y. Han, H. Li, C.Y. Ung, Z.Q. Tang, C.J. Zheng, Z.W. Cao, and Y.Z. Chen. 2007. Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol Immunol*. 44:514-520.

Dang, H.X., and C.B. Lawrence. 2014. Allerdictor: fast allergen prediction using text classification techniques. *Bioinformatics*.

de Castro, E., C.J. Sigrist, A. Gattiker, V. Bulliard, P.S. Langendijk-Genevaux, E. Gasteiger, A. Bairoch, and N. Hulo. 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*. 34:W362-365.

de la Monte, S., F. Castro, N.J. Bonilla, A. Gaskin de Urdaneta, and G.M. Hutchins. 1985. The systemic pathology of Venezuelan equine encephalitis virus infection in humans. *The American journal of tropical medicine and hygiene*. 34:194-202.

Delaney, B., J.D. Astwood, H. Cunny, R.E. Conn, C. Herouet-Guicheney, S. Macintosh, L.S. Meyer, L. Privalle, Y. Gao, J. Mattsson, M. Levine, and I.I.F.B.C.T.F.o.P. Safety. 2008. Evaluation of protein safety in the context of agricultural biotechnology. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association*. 46 Suppl 2:S71-97.

Dinkel, H., K. Van Roey, S. Michael, N.E. Davey, R.J. Weatheritt, D. Born, T. Speck, D. Kruger, G. Grebnev, M. Kuban, M. Strumillo, B. Uyar, A. Budd, B. Altenberg, M. Seiler, L.B. Chemes, J. Glavina, I.E. Sanchez, F. Diella, and T.J. Gibson. 2014. The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res*. 42:D259-266.

Dupuy, L.C., C.P. Locher, M. Paidhungat, M.J. Richards, C.M. Lind, R. Bakken, M.D. Parker, R.G. Whalen, and C.S. Schmaljohn. 2009. Directed molecular evolution improves the immunogenicity and protective efficacy of a Venezuelan equine encephalitis virus DNA vaccine. *Vaccine*. 27:4152-4160.

Edwards, R.J., N.E. Davey, and D.C. Shields. 2007. SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. *PLoS One*. 2:e967.

Egger, M., S. Mutschlechner, N. Wopfner, G. Gadermaier, P. Briza, and F. Ferreira. 2006. Pollen-food syndromes associated with weed pollinosis: an update from the molecular point of view. *Allergy*. 61:461-476.

Eswar, N., B. Webb, M.A. Marti-Renom, M.S. Madhusudhan, D. Eramian, M.Y. Shen, U. Pieper, and A. Sali. 2006. Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics*. Chapter 5:Unit 5 6.

Faisca, P.F., R.D. Travasso, T. Charters, A. Nunes, and M. Cieplak. 2010. The folding of knotted proteins: insights from lattice simulations. *Phys Biol*. 7:16009.

FAO/WHO. 2001. Evaluation of allergenicity of genetically modified foods. *In* Report of a joint FAO/WHO expert consultation.

FAO/WHO. 2003. Report of the fourth session of the codex ad hoc intergovernmental task force on foods derived from biotechnology.

FDA. 2004. Food Allergen Labeling and Consumer Protection Act of 2004 (Public Law 108-282, Title II).

Fiers, M.W., G.A. Kleter, H. Nijland, A.A. Peijnenburg, J.P. Nap, and R.C. van Ham. 2004. Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC bioinformatics*. 5:133.

Fine, D.L., B.A. Roberts, S.J. Terpening, J. Mott, D. Vasconcelos, and R.V. House. 2008. Neurovirulence evaluation of Venezuelan equine encephalitis (VEE) vaccine candidate V3526 in nonhuman primates. *Vaccine*. 26:3497-3506.

Finn, R.D., A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L. Sonnhammer, J. Tate, and M. Punta. 2014. Pfam: the protein families database. *Nucleic Acids Res*. 42:D222-230.

Flinn, A., and J. Hourihane. 2013. Allergic Reaction to Peanuts: Can We Predict Reaction Severity in the Wild? *Current Allergy and Asthma Reports*. 13:645-650.

Forman, J.R., and J. Clarke. 2007. Mechanical unfolding of proteins: insights into biology, structure and folding. *Curr Opin Struct Biol*. 17:58-66.

Fraczkiewicz, R.a.B., W. 1998. Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients for Macromolecules. *J. Comp. Chem*:319-333.

Galera-Prat, A., A. Gomez-Sicilia, A.F. Oberhauser, M. Cieplak, and M. Carrion-Vazquez. 2010. Understanding biology by stretching proteins: recent progress. *Curr Opin Struct Biol*. 20:63-69.

Gao, M., H. Lu, and K. Schulten. 2002a. Unfolding of titin domains studied by molecular dynamics simulations. *J Muscle Res Cell Motil*. 23:513-521.

Gao, M., M. Wilmanns, and K. Schulten. 2002b. Steered molecular dynamics studies of titin I1 domain unfolding. *Biophys J*. 83:3435-3445.

Garcia, T.I., A.F. Oberhauser, and W. Braun. 2009. Mechanical stability and differentially conserved physical-chemical properties of titin Ig-domains. *Proteins*. 75:706-718.

Gendel, S.M. 2002. Sequence analysis for assessing potential allergenicity. *Annals of the New York Academy of Sciences*. 964:87-98.

Gerardin, P., A. Fianu, D. Malvy, C. Mussard, K. Boussaid, O. Rollot, A. Michault, B.A. Gauzere, G. Breart, and F. Favier. 2011. Perceived morbidity and community burden after a Chikungunya outbreak: the TELECHIK survey, a population-based cohort study. *BMC medicine*. 9:5.

Gevers, D., R. Knight, J.F. Petrosino, K. Huang, A.L. McGuire, B.W. Birren, K.E. Nelson, O. White, B.A. Methe, and C. Huttenhower. 2012. The Human Microbiome Project: a community resource for the healthy human microbiome. *PLoS biology*. 10:e1001377.

Giltner, L.T., Shahan, M.S. 1933. The 1933 outbreak of infectious equine encephalomyelitis in the eastern states. *North Am. Vet.* 14:25-27.

Go, N., and H. Abe. 1981. Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation. *Biopolymers*. 20:991-1011.

Goodman, R.E. 2006. Practical and predictive bioinformatics methods for the identification of potentially cross-reactive protein matches. *Molecular nutrition & food research*. 50:655-660.

Goodman, R.E. 2008. Performing IgE serum testing due to bioinformatics matches in the allergenicity assessment of GM crops. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association*. 46 Suppl 10:S24-34.

Goodman, R.E., and J.N. Leach. 2004. Assessing the allergenicity of proteins introduced into genetically modified crops using specific human IgE assays. *Journal of AOAC International*. 87:1423-1432.

Goodman, R.E., and A.O. Tetteh. 2011. Suggested improvements for the allergenicity assessment of genetically modified plants used in foods. *Curr Allergy Asthma Rep*. 11:317-324.

Goodman, R.E., S. Vieths, H.A. Sampson, D. Hill, M. Ebisawa, S.L. Taylor, and R. van Ree. 2008. Allergenicity assessment of genetically modified crops--what makes sense? *Nature biotechnology*. 26:73-81.

Gough, J., K. Karplus, R. Hughey, and C. Chothia. 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol*. 313:903-919.

Gouy, M., S. Guindon, and O. Gascuel. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution*. 27:221-224.

Grandadam, M., V. Caro, S. Plumet, J.M. Thiberge, Y. Souares, A.B. Failloux, H.J. Tolou, M. Budelot, D. Cosserat, I. Leparc-Goffart, and P. Despres. 2011.

Chikungunya virus, southeastern France. *Emerging infectious diseases*. 17:910-913.

Granzier, H.L., and S. Labeit. 2004. The giant protein titin: a major player in myocardial mechanics, signaling, and disease. *Circulation research*. 94:284-295.

Griffith, I.J., J. Pollock, D.G. Klapper, B.L. Rogers, and A.K. Nault. 1991. Sequence polymorphism of Amb a I and Amb a II, the major allergens in Ambrosia artemisiifolia (short ragweed). *International archives of allergy and applied immunology*. 96:296-304.

Group, N.H.W., J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J.A. Schloss, V. Bonazzi, J.E. McEwen, K.A. Wetterstrand, C. Deal, C.C. Baker, V. Di Francesco, T.K. Howcroft, R.W. Karp, R.D. Lunsford, C.R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, A.R. Little, H. Peavy, C. Pontzer, M. Portnoy, M.H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer. 2009. The NIH Human Microbiome Project. *Genome research*. 19:2317-2323.

Grundy, J., S. Matthews, B. Bateman, T. Dean, and S.H. Arshad. 2002. Rising prevalence of allergy to peanut in children: Data from 2 sequential cohorts. *The Journal of allergy and clinical immunology*. 110:784-789.

Gupta, R.S., E.E. Springston, M.R. Warrier, B. Smith, R. Kumar, J. Pongracic, and J.L. Holl. 2011. The prevalence, severity, and distribution of childhood food allergy in the United States. *Pediatrics*. 128:e9-17.

Guzman, D.L., A. Randall, P. Baldi, and Z. Guan. 2010. Computational and single-molecule force studies of a macro domain protein reveal a key molecular determinant for mechanical stability. *Proc Natl Acad Sci U S A*. 107:1989-1994.

Hackman, P., A. Vihola, H. Haravuori, S. Marchand, J. Sarparanta, J. De Seze, S. Labeit, C. Witt, L. Peltonen, I. Richard, and B. Udd. 2002. Tibial muscular dystrophy is a titinopathy caused by mutations in TTN, the gene encoding the giant skeletal-muscle protein titin. *Am J Hum Genet*. 71:492-500.

Hart, M.K., K. Caswell-Stephan, R. Bakken, R. Tammariello, W. Pratt, N. Davis, R.E. Johnston, J. Smith, and K. Steele. 2000. Improved mucosal protection against Venezuelan equine encephalitis virus is induced by the molecularly defined, live-attenuated V3526 vaccine candidate. *Vaccine*. 18:3067-3075.

Hawley, R.J., and E.M. Eitzen, Jr. 2001. Biological weapons--a primer for microbiologists. *Annual review of microbiology*. 55:235-253.

Herman, R.A., P. Song, and A. Thirumalaiswamysekhar. 2009. Value of eight-amino-acid matches in predicting the allergenicity status of proteins: an empirical bioinformatic investigation. *Clinical and molecular allergy : CMA*. 7:9.

Hileman, R.E., A. Silvanovich, R.E. Goodman, E.A. Rice, G. Holleschak, J.D. Astwood, and S.L. Hefle. 2002. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *International archives of allergy and immunology*. 128:280-291.

Ho, B.K., and D.A. Agard. 2010. An improved strategy for generating forces in steered molecular dynamics: the mechanical unfolding of titin, e2lip3 and ubiquitin. *PLoS One*. 5.

Holm, L., and J. Park. 2000. DaliLite workbench for protein structure comparison. *Bioinformatics*. 16:566-567.

Hsin, J., J. Strumpfer, E.H. Lee, and K. Schulten. 2011. Molecular origin of the hierarchical elasticity of titin: simulation, experiment, and theory. *Annu Rev Biophys*. 40:187-203.

Hugouvieux-Cotte-Pattat, N., Condemine, G. and Shevchik, V. E. 2014. Bacterial pectate lyases, structural and functional diversity. *Environmental Microbiology Reports*.

Human Microbiome Project, C. 2012. A framework for human microbiome research. *Nature*. 486:215-221.

Hummer, G., and A. Szabo. 2003. Kinetics from nonequilibrium single-molecule pulling experiments. *Biophys J*. 85:5-15.

Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *Journal of molecular graphics*. 14:33-38, 27-38.

Hunt, A.R., S. Frederickson, T. Maruyama, J.T. Roehrig, and C.D. Blair. 2010. The first human epitope map of the alphaviral E1 and E2 proteins reveals a new E2 epitope with significant virus neutralizing activity. *PLoS Negl Trop Dis*. 4:e739.

Improta, S., A.S. Politou, and A. Pastore. 1996. Immunoglobulin-like modules from titin I-band: extensible components of muscle elasticity. *Structure*. 4:323-337.

ISAAA. 2010. Global status of commercialized Biotech/GM crops: executive summary, ISAAA brief no. 42. Ithaca: ISAAA.

Ivanciuc, O., T. Garcia, M. Torres, C.H. Schein, and W. Braun. 2009a. Characteristic motifs for families of allergenic proteins. *Mol Immunol*. 46:559-568.

Ivanciuc, O., T. Midoro-Horiuti, C.H. Schein, L. Xie, G.R. Hillman, R.M. Goldblum, and W. Braun. 2009b. The property distance index PD predicts peptides that cross-react with IgE antibodies. *Mol Immunol*. 46:873-883.

Ivanciuc, O., N. Oezguen, V.S. Mathura, C.H. Schein, Y. Xu, and W. Braun. 2004. Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Curr Med Chem*. 11:583-593.

Ivanciuc, O., C.H. Schein, and W. Braun. 2003. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res*. 31:359-362.

Ivanciuc, O., C.H. Schein, T. Garcia, N. Oezguen, S.S. Negi, and W. Braun. 2009c. Structural analysis of linear and conformational epitopes of allergens. *Regul Toxicol Pharmacol*. 54:S11-19.

Jenkins, J.A., S. Griffiths-Jones, P.R. Shewry, H. Breiteneder, and E.N. Mills. 2005. Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: an in silico analysis. *The Journal of allergy and clinical immunology*. 115:163-170.

Jiang, S., S. Wang, Y. Sun, Z. Zhou, and G. Wang. 2011. Molecular characterization of major allergens Ara h 1, 2, 3 in peanut seed. *Plant cell reports*. 30:1135-1143.

Kam, Y.W., W.W. Lee, D. Simarmata, R. Le Grand, H. Tolou, A. Merits, P. Roques, and L.F. Ng. 2014. Unique epitopes recognized by antibodies induced in Chikungunya virus-infected non-human primates: implications for the study of immunopathology and vaccine development. *PLoS One*. 9:e95647.

Kam, Y.W., F.M. Lum, T.H. Teo, W.W. Lee, D. Simarmata, S. Harjanto, C.L. Chua, Y.F. Chan, J.K. Wee, A. Chow, R.T. Lin, Y.S. Leo, R. Le Grand, I.C. Sam, J.C. Tong, P. Roques, K.H. Wiesmuller, L. Renia, O. Rotzschke, and L.F. Ng. 2012. Early neutralizing IgG response to Chikungunya virus in infected patients targets a dominant linear epitope on the E2 glycoprotein. *EMBO molecular medicine*. 4:330-343.

Kerr, P.J., S. Fitzgerald, G.W. Tregear, L. Dalgarno, and R.C. Weir. 1992. Characterization of a major neutralization domain of Ross river virus using anti-viral and anti-peptide antibodies. *Virology*. 187:338-342.

Kindt, T.J., R.A. Goldsby, B.A. Osborne, and J. Kuby. 2007. Kuby immunology. W.H. Freeman, New York. xxii, 574, A-531, G-512, AN-527, I-527 p. pp.

Kinney, R.M., G.J. Chang, K.R. Tsuchiya, J.M. Sneider, J.T. Roehrig, T.M. Woodward, and D.W. Trent. 1993. Attenuation of Venezuelan equine encephalitis virus strain TC-83 is encoded by the 5'-noncoding region and the E2 envelope glycoprotein. *J Virol*. 67:1269-1277.

Kleter, G.A., and A.A. Peijnenburg. 2002. Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE - binding linear epitopes of allergens. *BMC structural biology*. 2:8.

Kostyuchenko, V.A., J. Jakana, X. Liu, A.D. Haddow, M. Aung, S.C. Weaver, W. Chiu, and S.M. Lok. 2011. The structure of barmah forest virus as revealed by cryo-electron microscopy at a 6-angstrom resolution has detailed transmembrane protein architecture and interactions. *J Virol*. 85:9327-9333.

Kruger, M., and W.A. Linke. 2009. Titin-based mechanical signalling in normal and failing myocardium. *Journal of molecular and cellular cardiology*. 46:490-498.

Kruger, M., and W.A. Linke. 2011. The giant protein titin: a regulatory node that integrates myocyte signaling pathways. *J Biol Chem*. 286:9905-9912.

Kullback, S., and R.A. Leibler. 1951. On Information and Sufficiency. *Ann Math Stat*. 22:79-86.

Ladics, G.S., G.A. Bannon, A. Silvanovich, and R.F. Cressman. 2007. Comparison of conventional FASTA identity searches with the 80 amino acid sliding window FASTA search for the elucidation of potential identities to known allergens. *Molecular nutrition & food research*. 51:985-998.

Lanciotti, R.S., O.L. Kosoy, J.J. Laven, A.J. Panella, J.O. Velez, A.J. Lambert, and G.L. Campbell. 2007. Chikungunya virus in US travelers returning from India, 2006. *Emerging infectious diseases*. 13:764-767.

Lander, E.S., L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J.P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J.C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R.H. Waterston, R.K. Wilson, L.W. Hillier, J.D. McPherson, M.A. Marra, E.R. Mardis, L.A. Fulton, A.T. Chinwalla, K.H. Pepin, W.R. Gish, S.L. Chissoe, M.C. Wendl, K.D. Delehaunty, T.L. Miner, A. Delehaunty, J.B. Kramer, L.L. Cook, R.S. Fulton, D.L. Johnson, P.J. Minx, S.W. Clifton, T. Hawkins, E.

Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J.F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, et al. 2001. Initial sequencing and analysis of the human genome. *Nature*. 409:860-921.

Larkin, M.A., G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, and D.G. Higgins. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics*. 23:2947-2948.

Lee, E.H., J. Hsin, M. Sotomayor, G. Comellas, and K. Schulten. 2009. Discovery through the computational microscope. *Structure*. 17:1295-1306.

Leparc-Goffart, I., A. Nougairede, S. Cassadou, C. Prat, and X. de Lamballerie. 2014. Chikungunya in the Americas. *Lancet*. 383:514.

Li, H., M. Carrion-Vazquez, A.F. Oberhauser, P.E. Marszalek, and J.M. Fernandez. 2000. Point mutations alter the mechanical stability of immunoglobulin modules. *Nat Struct Biol*. 7:1117-1120.

Li, H., and J.M. Fernandez. 2003. Mechanical design of the first proximal Ig domain of human cardiac titin revealed by single molecule force spectroscopy. *J Mol Biol*. 334:75-86.

Li, H., W.A. Linke, A.F. Oberhauser, M. Carrion-Vazquez, J.G. Kerkvliet, H. Lu, P.E. Marszalek, and J.M. Fernandez. 2002. Reverse engineering of the giant muscle protein titin. *Nature*. 418:998-991002.

Li, H.B., Y. Cao, C. Lam, and M.J. Wang. 2006. Nonmechanical protein can have significant mechanical stability. *Angew Chem Int Edit*. 45:642-645.

Li, K.B., P. Issac, and A. Krishnan. 2004. Predicting allergenic proteins using wavelet transform. *Bioinformatics*. 20:2572-2578.

Li, L., J. Jose, Y. Xiang, R.J. Kuhn, and M.G. Rossmann. 2010. Structural changes of envelope proteins during alphavirus fusion. *Nature*. 468:705-708.

Li, W., and A. Godzik. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 22:1658-1659.

Linke, W.A. 2008. Sense and stretchability: the role of titin and titin-associated proteins in myocardial stress-sensing and mechanical dysfunction. *Cardiovascular research*. 77:637-648.

Linke, W.A., and N. Hamdani. 2014. Gigantic business: titin properties and function through thick and thin. *Circulation research*. 114:1052-1068.

Linke, W.A., and M. Kruger. 2010. The giant protein titin as an integrator of myocyte signaling pathways. *Physiology*. 25:186-198.

Lobigs, M., H.X. Zhao, and H. Garoff. 1990. Function of Semliki Forest virus E3 peptide in virus assembly: replacement of E3 with an artificial signal peptide abolishes spike heterodimerization and surface expression of E1. *J Virol*. 64:4346-4355.

Lord, R.D. 1974. History and geographic distribution of Venezuelan equine encephalitis. *Bulletin of the Pan American Health Organization*. 8:100-110.

Lu, H., B. Isralewitz, A. Krammer, V. Vogel, and K. Schulten. 1998. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys J*. 75:662-671.

Lu, H., and K. Schulten. 2000. The key event in force-induced unfolding of Titin's immunoglobulin domains. *Biophys J*. 79:51-65.

Lu, W., S.S. Negi, A.F. Oberhauser, and W. Braun. 2012. Engineering proteins with enhanced mechanical stability by force-specific sequence motifs. *Proteins*. 80:1308-1315.

Ludwig, G.V., M.J. Turell, P. Vogel, J.P. Kondig, W.K. Kell, J.F. Smith, and W.D. Pratt. 2001. Comparative neurovirulence of attenuated and non-attenuated strains of Venezuelan equine encephalitis virus in mice. *The American journal of tropical medicine and hygiene*. 64:49-55.

Magrane, M., and U. Consortium. 2011. UniProt Knowledgebase: a hub of integrated protein data. *Database : the journal of biological databases and curation*. 2011:bar009.

Marin-Rodriguez, M.C., J. Orchard, and G.B. Seymour. 2002. Pectate lyases, cell wall degradation and fruit softening. *Journal of experimental botany*. 53:2115-2119.

Marszalek, P.E., H. Lu, H. Li, M. Carrion-Vazquez, A.F. Oberhauser, K. Schulten, and J.M. Fernandez. 1999. Mechanical unfolding intermediates in titin modules. *Nature*. 402:100-103.

Martinez Barrio, A., D. Soeria-Atmadja, A. Nister, M.G. Gustafsson, U. Hammerling, and E. Bongcam-Rudloff. 2007. EVALLER: a web server for in silico assessment of potential protein allergenicity. *Nucleic Acids Res*. 35:W694-700.

Mathura, V.S., and W. Braun. 2001. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *Journal of Molecular Modeling*. 7:445-453.

Mathura, V.S., C.H. Schein, and W. Braun. 2003. Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases. *Bioinformatics*. 19:1381-1390.

Mavalankar, D., P. Shastri, T. Bandyopadhyay, J. Parmar, and K.V. Ramani. 2008. Increased mortality rate associated with chikungunya epidemic, Ahmedabad, India. *Emerging infectious diseases*. 14:412-415.

Mayans, O., J. Wuerges, S. Canela, M. Gautel, and M. Wilmanns. 2001. Structural evidence for a possible role of reversible disulphide bridge formation in the elasticity of the muscle protein titin. *Structure*. 9:331-340.

Metz, S.W., and G.P. Pijlman. 2011. Arbovirus vaccines; opportunities for the baculovirus-insect cell expression system. *J Invertebr Pathol*. 107 Suppl:S16-30.

Meyer, K.F., C.M. Haring, and B. Howitt. 1931. The Etiology of Epizootic Encephalomyelitis of Horses in the San Joaquin Valley, 1930. *Science*. 74:227-228.

Mi, T., J.C. Merlin, S. Deverasetty, M.R. Gryk, T.J. Bill, A.W. Brooks, L.Y. Lee, V. Rathnayake, C.A. Ross, D.P. Sargeant, C.L. Strong, P. Watts, S. Rajasekaran, and M.R. Schiller. 2012. Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic Acids Res*. 40:D252-260.

Midoro-Horiuti, T., R.M. Goldblum, A. Kurosky, T.G. Wood, C.H. Schein, and E.G. Brooks. 1999. Molecular cloning of the mountain cedar (Juniperus ashei) pollen major allergen, Jun a 1. *The Journal of allergy and clinical immunology*. 104:613-617.

Midoro-Horiuti, T., V. Mathura, C.H. Schein, W. Braun, S. Yu, M. Watanabe, J.C. Lee, E.G. Brooks, and R.M. Goldblum. 2003. Major linear IgE epitopes of mountain cedar pollen allergen Jun a 1 map to the pectate lyase catalytic site. *Mol Immunol*. 40:555-562.

Mittag, D., J. Akkerdaas, B.K. Ballmer-Weber, L. Vogel, M. Wensing, W.M. Becker, S.J. Koppelman, A.C. Knulst, A. Helbling, S.L. Hefle, R. Van Ree, and S. Vieths. 2004. Ara h 8, a Bet v 1-homologous allergen from peanut, is a major allergen in patients with combined birch pollen and peanut allergy. *The Journal of allergy and clinical immunology*. 114:1410-1417.

Mittag, D., V. Batori, P. Neudecker, R. Wiche, E.P. Friis, B.K. Ballmer-Weber, S. Vieths, and E.L. Roggen. 2006. A novel approach for investigation of specific and cross-reactive IgE epitopes on Bet v 1 and homologous food allergens in individual patients. *Mol Immunol*. 43:268-278.

Mittag, D., S. Vieths, L. Vogel, D. Wagner-Loew, A. Starke, P. Hunziker, W.M. Becker, and B.K. Ballmer-Weber. 2005. Birch pollen-related food allergy to legumes: identification and characterization of the Bet v 1 homologue in mungbean (Vigna radiata), Vig r 1. *Clin Exp Allergy*. 35:1049-1055.

Muh, H.C., J.C. Tong, and M.T. Tammi. 2009. AllerHunter: a SVM-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins. *PLoS One*. 4:e5861.

Mukhopadhyay, S., W. Zhang, S. Gabler, P.R. Chipman, E.G. Strauss, J.H. Strauss, T.S. Baker, R.J. Kuhn, and M.G. Rossmann. 2006. Mapping the structure and function of the E1 and E2 glycoproteins in alphaviruses. *Structure*. 14:63-73.

Murzin, A.G., S.E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 247:536-540.

Negi, S.S., and W. Braun. 2007. Statistical analysis of physical-chemical properties and prediction of protein-protein interfaces. *J Mol Model*. 13:1157-1167.

Negi, S.S., C.H. Schein, N. Oezguen, T.D. Power, and W. Braun. 2007. InterProSurf: a web server for predicting interacting sites on protein surfaces. *Bioinformatics*. 23:3397-3399.

Ng, S.P., K.S. Billings, T. Ohashi, M.D. Allen, R.B. Best, L.G. Randles, H.P. Erickson, and J. Clarke. 2007. Designing an extracellular matrix protein with enhanced mechanical stability. *Proc Natl Acad Sci U S A*. 104:9633-9637.

Ni, H., N.E. Yun, M.A. Zacks, S.C. Weaver, R.B. Tesh, A.P. da Rosa, A.M. Powers, I. Frolov, and S. Paessler. 2007. Recombinant alphaviruses are safe and useful serological diagnostic tools. *The American journal of tropical medicine and hygiene*. 76:774-781.

Noridah, O., Paranthaman, V., Nayar, S.K., Masliza, M., Ranjit, K., Norizah, I., Chem, Y.K., Mustafa, B., Kumarasamy, V., Chua, K.B. 2007. Outbreak of chikungunya due to virus of Central/East African genotype in Malaysia. *Medical Journal of Malaysia*. 62:323-328.

Oberhauser, A.F., and M. Carrion-Vazquez. 2008. Mechanical biochemistry of proteins one molecule at a time. *J Biol Chem*. 283:6617-6621.

Oezguen, N., S. Kumar, A. Hindupur, W. Braun, B.K. Muralidhara, and J.R. Halpert. 2008a. Identification and analysis of conserved sequence motifs in cytochrome P450 family 2. Functional and structural role of a motif 187RFDYKD192 in CYP2B enzymes. *J Biol Chem*. 283:21808-21816.

Oezguen, N., B. Zhou, S.S. Negi, O. Ivanciuc, C.H. Schein, G. Labesse, and W. Braun. 2008b. Comprehensive 3D-modeling of allergenic proteins and amino acid composition of potential conformational IgE epitopes. *Mol Immunol*. 45:3740-3747.

Paci, E., and M. Karplus. 1999. Forced unfolding of fibronectin type 3 modules: an analysis by biased molecular dynamics simulations. *J Mol Biol*. 288:441-459.

Paessler, S., R.Z. Fayzulin, M. Anishchenko, I.P. Greene, S.C. Weaver, and I. Frolov. 2003. Recombinant sindbis/Venezuelan equine encephalitis virus is highly attenuated and immunogenic. *J Virol*. 77:9278-9286.

Paessler, S., and S.C. Weaver. 2009. Vaccines for Venezuelan equine encephalitis. *Vaccine*. 27 Suppl 4:D80-85.

Pearson, W.R., and D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*. 85:2444-2448.

Pedersen, C.E., Jr., D.M. Robinson, and F.E. Cole, Jr. 1972. Isolation of the vaccine strain of Venezuelan equine encephalomyelitis virus from mosquitoes in Louisiana. *American journal of epidemiology*. 95:490-496.

Pfiffner, P., B.M. Stadler, C. Rasi, E. Scala, and A. Mari. 2012. Cross-reactions vs co-sensitization evaluated by in silico motifs and in vitro IgE microarray testing. *Allergy*. 67:210-216.

Pfuhl, M., S. Improta, A.S. Politou, and A. Pastore. 1997. When a module is also a domain: the role of the N terminus in the stability and the dynamics of immunoglobulin domains from titin. *J Mol Biol*. 265:242-256.

Phillips, J.C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, and K. Schulten. 2005. Scalable molecular dynamics with NAMD. *J Comput Chem*. 26:1781-1802.

Pittman, P.R., R.S. Makuch, J.A. Mangiafico, T.L. Cannon, P.H. Gibbs, and C.J. Peters. 1996. Long-term duration of detectable neutralizing antibodies after administration of live-attenuated VEE vaccine and following booster vaccination with inactivated VEE vaccine. *Vaccine*. 14:337-343.

Platts-Mills, T.A., E. Erwin, P. Heymann, and J. Woodfolk. 2005. Is the hygiene hypothesis still a viable explanation for the increased prevalence of asthma? *Allergy*. 60 Suppl 79:25-31.

Pleis, J.R., B.W. Ward, and J.W. Lucas. 2010. Summary health statistics for U.S. adults: National Health Interview Survey, 2009. *Vital and health statistics. Series 10, Data from the National Health Survey*:1-207.

Power, T.D., O. Ivanciuc, C.H. Schein, and W. Braun. 2013. Assessment of 3D models for allergen research. *Proteins*. 81:545-554.

Powers, A.M., and C.H. Logue. 2007. Changing patterns of chikungunya virus: re-emergence of a zoonotic arbovirus. *J Gen Virol*. 88:2363-2377.

Pratt, W.D., N.L. Davis, R.E. Johnston, and J.F. Smith. 2003. Genetically engineered, live attenuated vaccines for Venezuelan equine encephalitis: testing in animal models. *Vaccine*. 21:3854-3862.

Puchner, E.M., and H.E. Gaub. 2009. Force and function: probing proteins with AFM-based force spectroscopy. *Curr Opin Struct Biol*. 19:605-614.

Randall, R., F.D. Maurer, and J.E. Smadel. 1949. Immunization of laboratory workers with purified Venezuelan equine encephalomyelitis vaccine. *Journal of immunology*. 63:313-318.

Reed, D.S., M.G. Lackemeyer, N.L. Garza, S. Norris, S. Gamble, L.J. Sullivan, C.M. Lind, and J.L. Raymond. 2007. Severe encephalitis in cynomolgus macaques exposed to aerosolized Eastern equine encephalitis virus. *The Journal of infectious diseases*. 196:441-450.

Reed, D.S., C.M. Lind, L.J. Sullivan, W.D. Pratt, and M.D. Parker. 2004. Aerosol infection of cynomolgus macaques with enzootic strains of venezuelan equine encephalitis viruses. *The Journal of infectious diseases*. 189:1013-1017.

Relman, D.A., and S. Falkow. 2001. The meaning and impact of the human genome sequence for microbiology. *Trends in microbiology*. 9:206-208.

Rezza, G., L. Nicoletti, R. Angelini, R. Romi, A.C. Finarelli, M. Panning, P. Cordioli, C. Fortuna, S. Boros, F. Magurano, G. Silvi, P. Angelini, M. Dottori, M.G. Ciufolini, G.C. Majori, A. Cassone, and C.s. group. 2007. Infection with chikungunya virus in Italy: an outbreak in a temperate region. *Lancet*. 370:1840-1846.

Riaz, T., H.L. Hor, A. Krishnan, F. Tang, and K.B. Li. 2005. WebAllergen: a web server for predicting allergenic proteins. *Bioinformatics*. 21:2570-2571.

Rico, F., L. Gonzalez, I. Casuso, M. Puig-Vidal, and S. Scheuring. 2013. High-speed force spectroscopy unfolds titin at the velocity of molecular dynamics simulations. *Science*. 342:741-743.

Rinaldi, M., L. Harnack, C. Oberg, P. Schreiner, J. St Sauver, and L.L. Travis. 2012. Peanut allergy diagnoses among children residing in Olmsted County, Minnesota. *The Journal of allergy and clinical immunology*. 130:945-950.

Roehrig, J.T., and J.H. Mathews. 1985. The neutralization site on the E2 glycoprotein of Venezuelan equine encephalomyelitis (TC-83) virus is composed of multiple conformationally stable epitopes. *Virology*. 142:347-356.

Roussel, A., J. Lescar, M.C. Vaney, G. Wengler, and F.A. Rey. 2006. Structure and interactions at the viral surface of the envelope protein E1 of Semliki Forest virus. *Structure*. 14:75-86.

Sadler, D.P., E. Petrik, Y. Taniguchi, J.R. Pullen, M. Kawakami, S.E. Radford, and D.J. Brockwell. 2009. Identification of a mechanical rheostat in the hydrophobic core of protein L. *J Mol Biol*. 393:237-248.

Saha, S., and G.P. Raghava. 2006. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res*. 34:W202-209.

Salminen, A., J.M. Wahlberg, M. Lobigs, P. Liljestrom, and H. Garoff. 1992. Membrane fusion process of Semliki Forest virus. II: Cleavage-dependent reorganization of the spike protein complex controls virus entry. *The Journal of cell biology*. 116:349-357.

Satoh, M., M. Takahashi, T. Sakamoto, M. Hiroe, F. Marumo, and A. Kimura. 1999. Structural analysis of the titin gene in hypertrophic cardiomyopathy: identification of a novel disease gene. *Biochem Biophys Res Commun*. 262:411-417.

Savage, D.C. 1977. Microbial ecology of the gastrointestinal tract. *Annual review of microbiology*. 31:107-133.

Schaumann, T., W. Braun, and K. Wüthrich. 1990. The program FANTOM for energy refinement of polypeptides and proteins using a Newton – Raphson minimizer in torsion angle space. *Biopolymers*. 29:679-694.

Schein, C.H., O. Ivanciuc, and W. Braun. 2005a. Common physical-chemical properties correlate with similar structure of the IgE epitopes of peanut allergens. *Journal of agricultural and food chemistry*. 53:8752-8759.

Schein, C.H., O. Ivanciuc, and W. Braun. 2007. Bioinformatics approaches to classifying allergens and predicting cross-reactivity. *Immunol. Allerg. Clin. North Am.* 27:1-27.

Schein, C.H., N. Ozgun, T. Izumi, and W. Braun. 2002. Total sequence decomposition distinguishes functional modules, "molegos" in apurinic/apyrimidinic endonucleases. *BMC bioinformatics*. 3:37.

Schein, C.H., B. Zhou, N. Oezguen, V.S. Mathura, and W. Braun. 2005b. Molego-based definition of the architecture and specificity of metal-binding sites. *Proteins*. 58:200-210.

Seidman, J.G., and C. Seidman. 2001. The genetic basis for cardiomyopathy: from mutation identification to mechanistic paradigms. *Cell*. 104:557-567.

Sharma, D., Y. Cao, and H. Li. 2006. Engineering proteins with novel mechanical properties by recombination of protein fragments. *Angew Chem Int Ed Engl*. 45:5633-5638.

Sherman, M.B., J. Trujillo, I. Leahy, D. Razmus, R. Dehate, P. Lorcheim, M.A. Czarneski, D. Zimmerman, J.T. Newton, A.D. Haddow, and S.C. Weaver. 2013. Construction and organization of a BSL-3 cryo-electron microscopy laboratory at UTMB. *Journal of structural biology*. 181:223-233.

Shin, D.S., C.M. Compadre, S.J. Maleki, R.A. Kopper, H. Sampson, S.K. Huang, A.W. Burks, and G.A. Bannon. 1998. Biochemical and structural analysis of the IgE binding sites on ara h1, an abundant and highly allergenic peanut protein. *J Biol Chem*. 273:13753-13759.

Shreffler, W.G., D.A. Lencer, L. Bardina, and H.A. Sampson. 2005. IgE and IgG4 epitope mapping by microarray immunoassay reveals the diversity of immune response to the peanut allergen, Ara h 2. *The Journal of allergy and clinical immunology*. 116:893-899.

Sicherer, S.H. 2011. Epidemiology of food allergy. *The Journal of allergy and clinical immunology*. 127:594-602.

Sicherer, S.H., and H.A. Sampson. 2010. Food allergy. *The Journal of allergy and clinical immunology*. 125:S116-125.

Sigrist, C.J., E. de Castro, L. Cerutti, B.A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios. 2013. New and continuing developments at PROSITE. *Nucleic Acids Res*. 41:D344-347.

Sikora, M., J.I. Sulkowska, and M. Cieplak. 2009. Mechanical strength of 17,134 model proteins and cysteine slipknots. *PLoS computational biology*. 5:e1000547.

Sikora, M., J.I. Sulkowska, B.S. Witkowski, and M. Cieplak. 2011. BSDB: the biomolecule stretching database. *Nucleic Acids Res*. 39:D443-450.

Silvanovich, A., M.A. Nemeth, P. Song, R. Herman, L. Tagliani, and G.A. Bannon. 2006. The value of short amino acid sequence matches for prediction of protein allergenicity. *Toxicological sciences : an official journal of the Society of Toxicology*. 90:252-258.

Soeria-Atmadja, D., T. Lundell, M.G. Gustafsson, and U. Hammerling. 2006. Computational detection of allergenic proteins attains a new level of accuracy with in silico variable-length peptide extraction and machine learning. *Nucleic Acids Res*. 34:3779-3793.

Stacklies, W., M.C. Vega, M. Wilmanns, and F. Grater. 2009. Mechanical network in titin immunoglobulin from force distribution analysis. *PLoS computational biology*. 5:e1000306.

Stadler, M.B., and B.M. Stadler. 2003. Allergenicity prediction by protein sequence. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*. 17:1141-1143.

Stanley, J.S., N. King, A.W. Burks, S.K. Huang, H. Sampson, G. Cockrell, R.M. Helm, C.M. West, and G.A. Bannon. 1997. Identification and mutational analysis of the immunodominant IgE binding epitopes of the major peanut allergen Ara h 2. *Archives of biochemistry and biophysics*. 342:244-253.

Steward, A., J.L. Toca-Herrera, and J. Clarke. 2002. Versatile cloning system for construction of multimeric proteins for use in atomic force microscopy. *Protein Sci*. 11:2179-2183.

Strauss, E.G., D.S. Stec, A.L. Schmaljohn, and J.H. Strauss. 1991. Identification of antigenically important domains in the glycoproteins of Sindbis virus by analysis of antibody escape variants. *J Virol*. 65:4654-4664.

Sulkowska, J.I., and M. Cieplak. 2008. Selection of optimal variants of Go-like models of proteins through studies of stretching. *Biophys J*. 95:3174-3191.

Sulkowska, J.I., P. Sulkowski, P. Szymczak, and M. Cieplak. 2010. Untying knots in proteins. *J Am Chem Soc*. 132:13954-13956.

Sun, S., Y. Xiang, W. Akahata, H. Holdaway, P. Pal, X. Zhang, M.S. Diamond, G.J. Nabel, and M.G. Rossmann. 2013. Structural analyses at pseudo atomic resolution of Chikungunya virus and antibodies show mechanisms of neutralization. *eLife*. 2:e00435.

Sutton, L.S., and C.C. Brooke. 1954. Venezuelan equine encephalomyelitis due to vaccination in man. *Journal of the American Medical Association*. 155:1473-1476.

Tang, J., J. Jose, P. Chipman, W. Zhang, R.J. Kuhn, and T.S. Baker. 2011. Molecular links between the E2 envelope glycoprotein and nucleocapsid core in Sindbis virus. *J Mol Biol*. 414:442-459.

TenBroeck, C., Merrill, M.H. 1933. A serological difference between east-ern and western equine encephalomyelitis virus. *Proc. Soc. Exp. Biol. Med.* 31:217-220.

Tsai, C.J., S.L. Lin, H.J. Wolfson, and R. Nussinov. 1997. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci*. 6:53-64.

Tsetsarkin, K.A., R. Chen, M.B. Sherman, and S.C. Weaver. 2011. Chikungunya virus: evolution and genetic determinants of emergence. *Current opinion in virology*. 1:310-317.

Valbuena, A., J. Oroz, R. Hervas, A.M. Vera, D. Rodriguez, M. Menendez, J.I. Sulkowska, M. Cieplak, and M. Carrion-Vazquez. 2009. On the remarkable mechanostability of scaffoldins and the mechanical clamp motif. *Proc Natl Acad Sci U S A*. 106:13791-13796.

Vaney, M.C., S. Duquerroy, and F.A. Rey. 2013. Alphavirus structure: activation for entry at the target cell surface. *Current opinion in virology*. 3:151-158.

Volkova, E., E. Frolova, J.R. Darwin, N.L. Forrester, S.C. Weaver, and I. Frolov. 2008. IRES-dependent replication of Venezuelan equine encephalitis virus makes it highly attenuated and incapable of replicating in mosquito cells. *Virology*. 377:160-169.

Voss, J.E., M.C. Vaney, S. Duquerroy, C. Vonrhein, C. Girard-Blanc, E. Crublet, A. Thompson, G. Bricogne, and F.A. Rey. 2010. Glycoprotein organization of Chikungunya virus particles revealed by X-ray crystallography. *Nature*. 468:709-712.

Walton, T.E., O. Alvarez, Jr., R.M. Buckwalter, and K.M. Johnson. 1972. Experimental infection of horses with an attenuated Venezuelan equine encephalomyelitis vaccine (strain TC-83). *Infection and immunity*. 5:750-756.

Wang, J., D. Zhang, and J. Li. 2013. PREAL: prediction of allergenic protein by maximum Relevance Minimum Redundancy (mRMR) feature selection. *BMC systems biology*. 7 Suppl 5:S9.

Wang, K. 1996. Titin/connectin and nebulin: giant protein rulers of muscle structure and function. *Advances in biophysics*. 33:123-134.

Wang, Y., C.B. Harrison, K. Schulten, and J.A. McCammon. 2011. Implementation of Accelerated Molecular Dynamics in NAMD. *Comput Sci Discov*. 4.

WAO. 2011. WAO White Book on Allergy 2011-2012.

Weaver, S.C., C. Ferro, R. Barrera, J. Boshell, and J.C. Navarro. 2004. Venezuelan equine encephalitis. *Annu Rev Entomol*. 49:141-174.

Weaver, S.C., R. Salas, R. Rico-Hesse, G.V. Ludwig, M.S. Oberste, J. Boshell, and R.B. Tesh. 1996. Re-emergence of epidemic Venezuelan equine encephalomyelitis in South America. VEE Study Group. *Lancet*. 348:436-440.

Weaver, S.C., R. Winegar, I.D. Manger, and N.L. Forrester. 2012. Alphaviruses: population genetics and determinants of emergence. *Antiviral research*. 94:242-257.

Weinberg, E.G. 2011. The Wao White Book on Allergy 2011-2012. *Curr Allergy Clin Im*. 24:156-157.

Wiederstein, M., and M.J. Sippl. 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*. 35:W407-410.

Williams, P.M., S.B. Fowler, R.B. Best, J.L. Toca-Herrera, K.A. Scott, A. Steward, and J. Clarke. 2003. Hidden complexity in the mechanical properties of titin. *Nature*. 422:446-449.

Wing, R.A., J. Yamaguchi, S.K. Larabell, V.M. Ursin, and S. McCormick. 1990. Molecular and genetic characterization of two pollen-expressed genes that have sequence similarity to pectate lyases of the plant pathogen Erwinia. *Plant molecular biology*. 14:17-28.

Zacks, M.A., and S. Paessler. 2010. Encephalitic alphaviruses. *Veterinary microbiology*. 140:281-286.

Zhang, L., Y. Huang, Z. Zou, Y. He, X. Chen, and A. Tao. 2012. SORTALLER: predicting allergens using substantially optimized algorithm on allergen family featured peptides. *Bioinformatics*. 28:2178-2179.

Zhang, R., C.F. Hryc, Y. Cong, X. Liu, J. Jakana, R. Gorchakov, M.L. Baker, S.C. Weaver, and W. Chiu. 2011. 4.4 A cryo-EM structure of an enveloped alphavirus Venezuelan equine encephalitis virus. *The EMBO journal*. 30:3854-3863.